Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2011 *Paper* 80

A unified approach to non-negative matrix factorization and probabilistic latent semantic indexing

Karthik Devarajan* Guoli Wang†
Nader Ebrahimi‡

*Fox Chase Cancer Center, karthik.devarajan@fccc.edu
†wang.guoli2004@gmail.com
‡nader@math.niu.edu

A unified approach to non-negative matrix factorization and probabilistic latent semantic indexing

Karthik Devarajan, Guoli Wang, and Nader Ebrahimi

Abstract

Non-negative matrix factorization (NMF) by the multiplicative updates algorithm is a powerful machine learning method for decomposing a high-dimensional nonnegative matrix V into two matrices, W and H, each with nonnegative entries, $V \sim WH$. NMF has been shown to have a unique parts-based, sparse representation of the data. The nonnegativity constraints in NMF allow only additive combinations of the data which enables it to learn parts that have distinct physical representations in reality. In the last few years, NMF has been successfully applied in a variety of areas such as natural language processing, information retrieval, image processing, speech recognition and computational biology for the analysis and interpretation of large-scale data.

We present a generalized approach to NMF based on Renyi's divergence between two non-negative matrices related to the Poisson likelihood. Our approach unifies various competing models and provides a unique framework for NMF. Furthermore, we generalize the equivalence between NMF and probabilistic latent semantic indexing, a well-known method used in text mining and document clustering applications. We evaluate the performance of our method in the unsupervised setting using consensus clustering and demonstrate its applicability using real-life and simulated data.

A unified approach to non-negative matrix factorization and probabilistic latent semantic indexing

Karthik Devarajan, Division of Population Science, Fox Chase Cancer Center, Philadelphia, PA 19111

Guoli Wang, SRA International Inc., Fairfax, VA 22033

Nader Ebrahimi, Division of Statistics, Northern Illinois University, DeKalb, IL 60115

Abstract

Non-negative matrix factorization (NMF) by the multiplicative updates algorithm is a powerful machine learning method for decomposing a high-dimensional nonnegative matrix V into two matrices, W and H, each with nonnegative entries, $V \sim WH$. NMF has been shown to have a unique parts-based, sparse representation of the data. The nonnegativity constraints in NMF allow only additive combinations of the data which enables it to learn parts that have distinct physical representations in reality. In the last few years, NMF has been successfully applied in a variety of areas such as natural language processing, information retrieval, image processing, speech recognition and computational biology for the analysis and interpretation of large-scale data.

We present a generalized approach to NMF based on Renyi's divergence between two non-negative matrices related to the Poisson likelihood. Our approach unifies various competing models and provides a unique framework for NMF. Furthermore, we generalize the equivalence between NMF and probabilistic latent semantic indexing, a well-known method used in text mining and document clustering applications. We evaluate the performance of our method in the unsupervised setting using consensus clustering and demonstrate its applicability using real-life and simulated data.

Index Terms

nonnegative matrix factorization, Renyi's divergence, Kullback-Leibler divergence, probabilistic latent semantic indexing, consensus clustering, misclassification rate, text mining, document clustering, biomedical informatics, EM algorithm, λ-log-likelihood, sparse, high-performance computing, message-passing interface

I. Introduction

Nonnegative matrix factorization (NMF) was introduced as an unsupervised parts-based learning paradigm in which a nonnegative matrix V is decomposed into two nonnegative matrices, W and H, such that $V \sim WH$, by a multiplicative updates algorithm [31,32]. In the past decade, NMF has been widely used in a variety of areas including natural language processing such as text mining and document clustering, information retrieval, image processing and facial pattern recognition, sparse coding, speech recognition, video summarization and Internet research. More recently, this approach has found its way into the domain of computational biology, particularly in the analysis and interpretation of high-throughput biological data. For a complete review of the applications of NMF, the interested reader is referred to [11] and references therein.

References [31,32] outlined algorithms for NMF based on Kullback-Leibler divergence and Euclidean distance. These are related to the Poisson and Gaussian likelihoods, respectively, between two nonnegative matrices. They applied it to text mining and facial pattern recognition. Since its introduction, several variants of their algorithm have been proposed in the literature. These include, but are not limited to, references [8,12,21,22,23,28,33,39,40,46,47,49]. In previous work [12], we generalized NMF based on Renyi's divergence between two non-negative matrices, also related to the Poisson likelihood [43]. Renyi's divergence is indexed by a parameter γ and represents a continuum of distance measures based on the choice of this parameter. In this paper, we develop a unique framework for NMF and provide a rigorous convergence proof of our generalized algorithm based on Renyi's divergence. Our approach includes several well-known distance measures as special cases. In addition, we generalize the relationship between NMF and probabilistic latent semantic indexing (PLSI), a commonly used method for text mining and document clustering applications. We show that the equivalence between these methods is embedded within our broader framework as a special case.

We demonstrate the utility of our generalized approach to NMF in unsupervised clustering. To that end, we use consensus clustering to quantitatively evaluate the homogeneity and accuracy of clustering. We illustrate the applicability of our methods to text mining and document clustering using several real-life and simulated data sets. The extension of our methods to other areas of application is straightforward.

This paper is organized as follows. Section 2 gives an overview of the fundamental concepts and provides an extensive discussion of Renyi's divergence and related distance measures. Section 3 explores the applicability of these measures in the context of NMF and provides update rules based on our generalized measure. In section 4, we generalize the equivalence between NMF and PLSI. In section 5, we describe the quantitative evaluation of clustering based on our approach and in section 6, we illustrate our methods in detail by applying it to a variety of real-life and simulated document clustering data sets. Section 7 provides a discussion and concluding remarks. Detailed proofs of convergence for our optimization algorithm are relegated to the Appendix.

II. A Generalized Divergence Measure

Consider the problem of discriminating between two probability models F and G for a random prospect X that ranges over the space S. Given an observation $X = x$, Bayes theorem relates

the likelihood ratio to the prior and posterior odds in favor of F as follows:

$$\log \frac{f(x)}{g(x)} = \log \frac{P(F|x)}{P(G|x)} - \log \frac{P(F)}{P(G)}, \tag{2.1}$$

where f and g are probability density (mass) functions, and $P(\cdot)$ and $P(\cdot|x)$ denote the prior and posterior probabilities of the model, respectively. As the difference between the posterior and prior log-odds, the logarithm of the likelihood ratio $\log \left[\frac{f(x)}{g(x)}\right]$ quantifies the information in $X = x$ in favor of F against G.

Suppose that x is not given and there is not specific information on the whereabouts of x, other than $x \epsilon S$, then the mean observation per x from F for the discrimination information between F and G is

$$K(f:g) = \int \left(\log \frac{f(x)}{g(x)}\right) dF(x), \tag{2.2}$$

given that F is absolutely continuous with respect to G. The discrimination information function (2.2) introduced in [29,30] is a measure for comparing two distributions. This is referred to as the Kullback-Leibler (KL) divergence. See [15] for properties of this measure.

Renyi's divergence, which is referred to as the information divergence of order γ between two distributions F and G, is defined by

$$R_\gamma(f:g) = \frac{1}{\gamma - 1} \log \int \left(\frac{f(x)}{g(x)}\right)^{\gamma - 1} dF(x), \tag{2.3}$$

where $\gamma \neq 1$ [43]. Various well-known distance measures arise from Renyi's divergence as special cases. For instance, in the limit $\gamma \to 1$, Renyi's divergence becomes the KL divergence $K(f:g)$. Information divergence of order $\gamma, R_\gamma(f:g)$ is also symmetric for the case $\gamma = \frac{1}{2}$, i.e., $R_{\frac{1}{2}}(f:g) = R_{\frac{1}{2}}(g:f)$. An important feature of Renyi's divergence is that if $Y = T(X)$ is a nonsingular transformation, then for any γ, $R_\gamma(f_X : g_X) = R_\gamma(f_Y : g_Y)$. That is, our measure is invariant under any nonsingular transformation on the original data.

For two Poisson random variables with parameters m_1 and m_2, i.e., $f(x) = \frac{e^{-m_1} m_1^x}{x!}$ and $g(x) = \frac{e^{-m_2} m_2^x}{x!}$, one can easily show that

$$\begin{aligned} R_\gamma(f:g) &= \frac{1}{\gamma - 1} \log \sum_{x=0}^{\infty} \left[\frac{e^{-m_1} m_1^x}{x!}\right]^\gamma \left[\frac{e^{-m_2} m_2^x}{x!}\right]^{1-\gamma} \\ &= \frac{1}{\gamma - 1} \left(-\gamma m_1 - (1-\gamma) m_2 + m_1^\gamma m_2^{1-\gamma}\right). \end{aligned} \tag{2.4}$$

As mentioned above, for the limiting case $\gamma \to 1$, Renyi's divergence becomes KL divergence

$$K(f:g) = m_1 \log\left(\frac{m_1}{m_2}\right) - m_1 + m_2. \tag{2.5}$$

In the special case that $\gamma = \frac{1}{2}$,

$$R_{\frac{1}{2}}(f:g) = R_{\frac{1}{2}}(g:f) = \left(\sqrt{m}_1 - \sqrt{m}_2\right)^2. \tag{2.6}$$

This is the well-known Bhattacharya distance and is a symmetric measure [17]. For this case, it is also the logarithm of the squared Matusita or Hellinger distance [36]. If $\gamma = 2$,

$$R_2(f:g) = \frac{(m_1 - m_2)^2}{m_2}, \tag{2.7}$$

which is the Pearson chi-squared estimator. For the case $\gamma = -1$, we obtain the modified chi-squared estimator due to [38]. And for $\gamma = \frac{5}{3}$, we obtain the Cressie-Read distance estimator [5].

Our motivation for a generalized approach to NMF using the Poisson likelihood is based on the power-divergence family of statistics [1]. In this context, it is given by

$$\phi_\lambda(m_1, m_2) = \frac{2}{\lambda(\lambda+1)} m_1 \left[\left(\frac{m_1}{m_2}\right)^\lambda - 1\right] \tag{2.8}$$

for $\lambda \neq -1$ and $\lambda \neq 0$. This family of measures and its variants have been extensively studied in the statistical literature in the context of discrete multivariate data analysis [5,6,42]. It is straightforward to obtain one measure from the other and all the special cases outlined above via reparametrizations. For example in (2.8), $\lambda \to 0$ corresponds to $\gamma \to 1$ in (2.4). Similarly, $\lambda = -\frac{1}{2}, -2$ and 1 correspond to $\gamma = \frac{1}{2}, -1$ and 2, respectively in (2.4). This generalization unifies various competing models into a unique framework for various applications in high-dimensional data analysis using NMF. In the next section, we will revisit this topic and discuss them in further detail in the context of NMF.

III. METHODS

Text mining and document clustering are concerned with the recognition of patterns or similarities in natural language text. Consider a corpus of documents that is summarized as a $p \times n$ matrix V in which the rows represent the terms in the vocabulary and the columns correspond to the documents in the corpus. The entries of V denote the frequencies of words in each document. In document clustering studies, the number of terms p is typically in the thousands and the number

of documents n is typically in the hundreds. The objective is to identify subsets of semantic categories and to cluster the documents based on their association with these categories. To this end, we propose to find a small number of metaterms, each defined as a nonnegative linear combination of the p terms. This is accomplished via a decomposition of the frequency matrix V into two matrices with nonnegative entries, $V \sim WH$, where W has size $p \times k$, with each of k columns defining a metaterm and the matrix H has size $k \times n$, with each of n columns representing the metaterm frequency pattern of the corresponding document. The rank k of the factorization is chosen so that $(n + p)k < np$. Here, the entry w_{ia} in the matrix W is the coefficient of term i in metaterm a and the entry h_{aj} in the matrix H quantifies the influence of metaterm a in sample j.

The nonnegativity constraints in NMF are compatible with the intuitive notion of combining parts to form a whole, i.e., they provide a parts based representation of the data. This is in contrast to a holistic representation of the data provided by VQ and the distributed representation provided by PCA [31]. The metaterms and the metaterm frequency patterns have a sparse representation, potentially representing local hidden variables or clusters. In the context of text mining, these clusters are subgroups of terms that co-occur in subgroups of documents. The perception of the whole is simply a combination of the parts represented by these basis vectors. Since the data are presented as frequency of occurrence of terms for each document, NMF provides a more natural representation of the metaterms and metaterm frequency patterns. Unlike PCA and VQ, the nonnegative coefficients in each metaterm are easily interpretable as the relative contribution of terms. In this setting, NMF has been shown to be superior to PCA (see [46] and references therein). Interestingly, the Poisson likelihood approach to NMF due to [31] owes its origin to an application in text mining involving count data. Another useful feature of NMF is that it does not force hierarchy into the data structure like some methods. However, it does identify a hierarchy when present [3]. For a more thorough discussion of the interpretation of the factorization and the nonnegativity constraints in NMF, the interested reader is referred to [11].

In this paper, our focus will be on clustering documents. However, we note that there is a dual view of the decomposition $V \sim WH$. This is achieved simply by taking the transpose $V' \sim (WH)' = H'W'$. A notable example of such an application is in biomedical informatics, which involves text mining and document clustering of biomedical literature. Reference [4] describes the application of NMF to create literature profiles from a corpus of documents relevant

to large sets of genes and proteins using common semantic features extracted from the corpus. Genes are then represented as additive linear combinations of the semantic features which can be further used for studying their functional associations. Using NMF, existing information about the biological entities under study can thus be utilized in establishing putative relationships among subsets of genes and proteins that characterize a subset of the data.

In order to find an approximate factorization for the matrix V, we first need to define functions that quantify the quality of the approximation. In general, such a function can be constructed using some measure of distance between any two nonnegative matrices, say A and B. Examples of such measures include the Euclidean distance and KL divergence. The latter can be derived based on reconstruction of an image represented by the matrix A from the matrix B by the addition of Poisson noise, i.e,

$$A = B + \epsilon \tag{3.1}$$

where ϵ is a Poisson random variable. This formulation was originally described in [31] for text mining applications involving count data as well as for facial pattern recognition.

We generalize this approach by using Renyi's divergence $R_\gamma(f : g)$ related to the Poisson likelihood of generating A from B, as described in section 2. Specifically, using (2.4), our measure is

$$D^*_\gamma(A||B) = \frac{1}{\gamma - 1} \sum_{i,j} \left[A_{ij}^\gamma B_{ij}^{1-\gamma} - \gamma A_{ij} - (1-\gamma) B_{ij} \right]. \tag{3.2}$$

In (2.4), if $\gamma \to 1$, then D^*_γ in (3.2) becomes KL divergence $K(A : B) = \sum_{i,j} A_{ij} \log \left(\frac{A_{ij}}{B_{ij}} \right) - A_{ij} + B_{ij}$. This coincides with the measure proposed by [31]. In fact, we can generalize the measure in (3.2) by re-defining it as

$$D^*_\gamma(A||B) = \frac{1}{\gamma(\gamma - 1)} \sum_{i,j} \left[A_{ij}^\gamma B_{ij}^{1-\gamma} - \gamma A_{ij} - (1-\gamma) B_{ij} \right] \tag{3.3}$$

where $\gamma \neq 1$ and $\gamma \neq 0$. This includes Renyi's divergence as defined by (2.4) and its special cases. Moreover, if $\gamma \to 0$, we obtain the dual KL divergence [29]

$$K(B : A) \;\; = \;\; \sum_{i,j} B_{ij} \log \left(\frac{B_{ij}}{A_{ij}} \right) - B_{ij} + A_{ij}. \tag{3.4}$$

For small values of γ, (3.4) provides a reasonable approximation of (3.3). For $\gamma \neq 1$ and $\gamma \neq 0$, one can ignore $\frac{1}{\gamma(\gamma-1)}$ in (3.3) and define the function

$$D_\gamma(A||B) = \begin{cases} \sum_{i,j} A_{ij}^\gamma B_{ij}^{1-\gamma} - \gamma A_{ij} - (1-\gamma)B_{ij}, \ \gamma > 1 \\ \\ \sum_{i,j} \gamma A_{ij} + (1-\gamma)B_{ij} - A_{ij}^\gamma B_{ij}^{1-\gamma}, \ 0 < \gamma < 1. \end{cases} \tag{3.5}$$

Thus, for any information measure which is proportional to Renyi's divergence, we obtain equation (3.5). Similarly for $\gamma \neq 1$, one can ignore $\frac{1}{\gamma-1}$ in (3.2) and define the function

$$D_\gamma(A||B) = \sum_{i,j} -\gamma A_{ij} - (1-\gamma)B_{ij} + A_{ij}^\gamma B_{ij}^{1-\gamma}, \ \gamma < 0. \tag{3.6}$$

The Euclidean distance (ED) between two nonnegative matrices A and B is simply given by $D(A||B) = \sum_{i,j} [A_{ij} - B_{ij}]^2$. It is interesting to note that the KL divergence between two normal random variables with means μ_1 and μ_2 and (equal) variance σ is simply the well-known ED given by $\frac{1}{2}(\mu_1 - \mu_2)^2$. While ED is not directly a member of the class of distance measures defined by Renyi's divergence, it is equivalent to the measure obtained by invoking the transformation invariant property of Renyi's divergence and applying it to $[{A_{ij}}^2]$ for the case $\gamma = \frac{1}{2}$. In this sense, ED can be considered to be a part of this family of distance measures. In the case of non-normal data such as those arising in text mining and document clustering applications, Renyi's divergence is a flexible choice in decomposing a document frequency matrix.

For a given document frequency matrix V, we now formally consider a method for finding nonnegative matrices W and H such that $V \approx WH$. In our setup, this is equivalent to minimizing $D_\gamma(V||WH)$ in (3.3) with respect to W and H, subject to the constraints $W, H \geq 0$. In this formulation, we observe that for a given γ, $D_\gamma(V||WH)$ is not convex in both variables (V and WH) together. Hence, the algorithm will only converge to a local minima. There are many techniques such as gradient descent and conjugate gradient from numerical optimization that can be applied to find the minima. In this paper, we use multiplicative update rules, similar to that in [32]. For a given γ, we will start with random initial values for W and H and iterate until convergence, i.e, iterate until $|D_\gamma^{(i)}(V||WH) - D_\gamma^{(i-1)}(V||WH)| < \delta$ where δ is a pre-specified threshold between 0 and 1 and i denotes the iteration number.

Theorem 1: For $\gamma > 0$, the measure $D_\gamma(V||WH)$ is non-increasing under the multiplicative update rules for W and H given by

$$H_{ak}^{t+1} = H_{ak}^{t} \left(\frac{\sum_i \left(\frac{V_{ik}}{\sum_b W_{ib} H_{bk}^t} \right)^\gamma W_{ia}}{\sum_i W_{ia}} \right)^{1/\gamma}$$

and

$$W_{ia}^{t+1} = W_{ia}^{t} \left(\frac{\sum_i \left(\frac{V_{ik}}{\sum_b W_{ib}^t H_{bk}} \right)^\gamma H_{ak}}{\sum_k H_{ak}} \right)^{1/\gamma} .$$

This measure is also invariant under these updates if and only if W and H are at a stationary point of the divergence.

Proof: See Appendix I.

IV. Equivalence of NMF and PLSI: A Generalization

Probabilistic latent semantic indexing (PLSI) is a method for modeling co-occurrence data arising in natural language processing such as text mining and document clustering. It is based on a statistical latent class model called the *aspect model* for the analysis of count data [19]. PLSI employs the likelihood principle and results in a factor representation of the data such as in NMF, thereby defining a proper generative model of the data.

Consider the corpus of documents summarized as a $p \times n$ co-occurrence matrix V described in section 3. The rows of V represent the terms in the vocabulary and the columns represent the documents in the corpus. The aspect model associates an unobserved class variable $k \in \{1, 2, ..., K\}$ (where $K < n$) with each occurrence of a term $i \in \{1, 2, ..., p\}$ in a document $j \in \{1, 2, ..., n\}$ via the following mixture model

$$P_{ij} = P(j) \sum_{k=1}^{K} P(i|k) P(k|j) \tag{4.1}$$

$$= \sum_{k=1}^{K} P(k) P(i|k) P(j|k) \tag{4.2}$$

where P_{ij} is the joint probability of generating the observation pair (i, j), $P(j)$ and $P(k|j)$ represent the probabilities of choosing document j and latent class k, respectively, and $P(i|k)$

is the probability of generating term i. The equivalence of terms on the right hand side of (4.1) and (4.2) can be shown using Bayes' theorem. The probabilities in (4.2) sum to unity, i.e.,

$$\sum_{ij} P_{ij} = \sum_{i=1}^{p} P(i|k) = \sum_{j=1}^{n} P(j/k) = \sum_{k=1}^{K} P(k) = 1 \tag{4.3}$$

Let v_{ij} denote the frequency of occurrence of term i in document j, i.e., the $(i,j)^{th}$ entry of the matrix V. In the context of NMF, v_{ij} has a Poisson distribution with mean μ_{ij} and the v_{ij}s are independent. Hence, the probability of $\{v_{ij}\}, i = 1, ..., p, j = 1, ..., n$ (or the likelihood) is

$$\prod_{ij} P(v_{ij}) = e^{-\sum_{ij} \mu_{ij}} \prod_{ij} \frac{{\mu_{ij}}^{v_{ij}}}{v_{ij}!} \tag{4.4}$$

and the log-likelihood can be shown to be equivalent to $\sum_{ij} \left\{ -v_{ij} \log \left(\frac{v_{ij}}{\mu_{ij}} \right) - \mu_{ij} + v_{ij} \right\}$.

In PLSI, we normalize the term frequencies by conditioning on their sum such that $\sum_{ij} v_{ij} = v$ where v is fixed. This is in contrast to NMF where $\sum_{ij} v_{ij}$ is random, rather than fixed, due to Poisson sampling. The normalized term frequencies v_{ij} are neither independent nor Poisson distributed. It can be shown that the probability of $\{v_{ij}\}, i = 1, ..., p, j = 1, ..., n$ conditional on this sum is

$$\prod_{ij} P(v_{ij} | \sum_{ij} v_{ij} = v) = \left(\frac{v!}{\prod_{ij} v_{ij}!} \right) \prod_{ij} {P_{ij}}^{v_{ij}} \tag{4.5}$$

where $P_{ij} = \frac{\mu_{ij}}{\sum_{ij} \mu_{ij}}$ [1]. This is based on multinomial sampling and represents the likelihood for PLSI. Typically, the term frequencies are normalized by re-scaling to their sum such that $v_{ij} \leftarrow \frac{v_{ij}}{\sum_{ij} v_{ij}}$ and $\sum_{ij} v_{ij} = 1$. Using (4.5), the log-likelihood for PLSI is equivalent to

$$\mathcal{L} = \sum_{j=1}^{n} \sum_{i=1}^{p} v_{ij} \log P_{ij} \tag{4.6}$$

[19]. We now generalize the likelihood for PLSI (4.6) in the following lemma.

a) **Lemma:** The log-likelihood for PLSI (4.6) is a member of the family of λ-log-likelihoods given by

$$\mathcal{L} = -\frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{n} \sum_{i=1}^{p} \left\{ v_{ij} \left[\left(\frac{v_{ij}}{P_{ij}} \right)^{\lambda} - 1 \right] - \lambda(v_{ij} - P_{ij}) \right\} \tag{4.7}$$

where $\lambda \neq 0$ and $\lambda \neq 1$.

Proof: Without loss of generality, we re-write the log-likelihood (4.6) by adding a constant term

as follows.

$$\begin{aligned}\mathcal{L} &= \sum_{j=1}^{n}\sum_{i=1}^{p} v_{ij}\log P_{ij} - v_{ij}\log v_{ij} \\ &= \sum_{j=1}^{n}\sum_{i=1}^{p} -v_{ij}\log\left(\frac{v_{ij}}{P_{ij}}\right) \end{aligned} \tag{4.8}$$

Using the Box-Cox family of transformations [2], we can generalize it as

$$\mathcal{L}_\lambda = \sum_{j=1}^{n}\sum_{i=1}^{p} -\frac{v_{ij}}{\lambda}\left[\left(\frac{v_{ij}}{P_{ij}}\right)^\lambda - 1\right] \tag{4.9}$$

where $\lambda \neq 0$. In the limit $\lambda \to 0$, we obtain the log-likelihood given in (4.6). This is similar in principle to the α-log-likelihood approach outlined in [35]. Since $\sum_{ij} v_{ij} = \sum_{ij} P_{ij} = 1$, we have $\sum_{ij}(v_{ij} - P_{ij}) = 0$. Adding this term to $\mathcal{L}_\lambda$ in (4.9) and multiplying throughout by the constant $\dfrac{2}{\lambda+1}$ does not alter the meaning and interpretation of $\mathcal{L}_\lambda$, and results in

$$\begin{aligned}\mathcal{L}_\lambda &= \sum_{j=1}^{n}\sum_{i=1}^{p} -\frac{v_{ij}}{\lambda}\left[\left(\frac{v_{ij}}{P_{ij}}\right)^\lambda - 1\right] \\ &= \frac{2}{\lambda+1}\sum_{j=1}^{n}\sum_{i=1}^{p}\left\{-\frac{v_{ij}}{\lambda}\left[\left(\frac{v_{ij}}{P_{ij}}\right)^\lambda - 1\right] + (v_{ij} - P_{ij})\right\} \end{aligned} \tag{4.10}$$

We refer to $\mathcal{L}_\lambda = \mathcal{L}_\lambda(V, P)$ in (4.10) as the λ-log-likelihood where $V = [v_{ij}]$, $P = [P_{ij}]$, $\lambda \neq -1$ and $\lambda \neq 0$. Maximizing $\mathcal{L}_\lambda(V, P)$ is equivalent to minimizing $-\mathcal{L}_\lambda(V, P)$ where

$$-\mathcal{L}_\lambda(V, P) = \frac{2}{\lambda(\lambda+1)}\sum_{j=1}^{n}\sum_{i=1}^{p}\left\{v_{ij}\left[\left(\frac{v_{ij}}{P_{ij}}\right)^\lambda - 1\right] - \lambda(v_{ij} - P_{ij})\right\} \tag{4.11}$$

where $\lambda \neq 0$ and $\lambda \neq -1$. □

Using KL divergence, references [13,14] showed equivalence between NMF and PLSI. In the following theorem, we show that this equivalence is embedded within our broader framework as a special case and therefore generalizes the relationship between NMF and PLSI.

Theorem 2: Renyi's divergence, $D^*_\gamma(A||B)$, between two non-negative matrices A and B in (3.3) is equivalent to the negative λ-log-likelihood $-\mathcal{L}_\lambda(A, B)$ given by (4.11)), and therefore generalizes the equivalence between NMF and PLSI.

Proof: In the context of NMF, the power-divergence family of statistics given in (2.8) can be re-written as

$$\phi_\lambda(A, B) = \frac{2}{\lambda(\lambda+1)}\sum_{i,j} A_{ij}\left[\left(\frac{A_{ij}}{B_{ij}}\right)^\lambda - 1\right] - \lambda(A_{ij} - B_{ij}) \tag{4.12}$$

for $\lambda \neq -1$ and $\lambda \neq 0$ since $\sum_{i,j}(A_{ij} - B_{ij}) = 0$. Note that $\phi_\lambda(A, B) = -\mathcal{L}_\lambda(A, B)$, and if we reparametrize (4.12) such that $A_{ij} = \frac{\gamma}{2}\tilde{A}_{ij}$, $B_{ij} = \frac{\gamma}{2}\tilde{B}_{ij}$ and $\lambda = \gamma - 1$, we obtain the quantity $D^*_\gamma(\tilde{A}||\tilde{B})$ defined in (3.3). Hence the negative λ-log-likelihood is equivalent to Renyi's divergence between the matrices A and B.

In the limit $\lambda \to 0$, we obtain

$$-\mathcal{L}_{\lambda\to 0}(A, B) = \sum_{j=1}^{n}\sum_{i=1}^{p} A_{ij}\log\left(\frac{A_{ij}}{B_{ij}}\right) - A_{ij} + B_{ij}, \tag{4.13}$$

the KL divergence between A and B. The case $\lambda = 1$ yields

$$-\mathcal{L}_{\lambda=1} = \sum_{j=1}^{n}\sum_{i=1}^{p}\frac{(A_{ij} - B_{ij})^2}{B_{ij}}, \tag{4.14}$$

the Pearson chi-squared statistic. If we reparametrize (4.12) such that $A_{ij} = \tilde{A}^2_{ij}$, $B_{ij} = \tilde{B}^2_{ij}$ for $\lambda = -\frac{1}{2}$, we obtain the Euclidean distance $\sum_{j=1}^{n}\sum_{i=1}^{p}(A_{ij} - B_{ij})^2$ between A and B. The equivalence between Renyi's divergence and the negative λ-log-likelihood thus generalizes the equivalence between NMF and PLSI. □

We note that [13] provided a first order approximation to KL divergence (4.13) using the Pearson chi-squared statistic (4.14) and also pointed out the relationship between Pearson chi-squared and Euclidean distance. Our unified representation elucidates the relationship between these measures whereby each measure is obtained as a special case for different choices of λ.

V. Quantitative Evaluation of Clustering

In this section, we describe the implementation of our NMF algorithm and quantitatively evaluate its performance in grouping n documents into homogeneous classes based on the frequency of occurrence of p terms. The NMF algorithm may not converge to the same solution on each run due to the random nature of initial conditions. We exploited this feature to evaluate the consistency of its performance and to quantify the clustering accuracy for a benchmark data set where the true number of classes k is known. The algorithm is applied multiple times with random initial starting values for W and H; and it groups the documents into k clusters, where k is the pre-specified rank of the factorization.

In order to assess whether a given γ provides a meaningful decomposition of the data for a fixed (known) number of classes $k = K$, we applied consensus clustering to evaluate the clustering accuracy of the factorization. Consensus clustering [3,37] evaluates the performance

of any unsupervised clustering algorithm based on resampling methods. In our case, the stochastic nature of initial conditions in the NMF algorithm is utilized in the evaluation process. In this approach, class membership for each document is determined based on the highest metaterm frequency profile. Each run of the algorithm results in an $n \times n$ connectivity matrix C with an entry of 1 if documents i and j cluster together and 0 otherwise, where $i, j = 1, , n$. The consensus matrix $\bar{C}$ is simply the average connectivity matrix obtained over N runs of the algorithm. Final document assignments are based on the re-ordered consensus matrix obtained by hierarchical clustering (HC) using average linkage. In our studies [9,10], we found the performance of the method to be consistent across multiple runs and, in general, 50-200 runs were sufficient to provide stability to the clustering. For a rank K factorization, we applied consensus clustering to different choices of γ, and evaluated the clustering accuracy for each γ based on the measures described in the next section.

A. Measures for Evaluating Clustering Accuracy

We utilize four measures for evaluating clustering accuracy by combining the information across multiple runs of the NMF algorithm. These are the misclassification rate, normalized mutual information, adjusted Rand index and the cophenetic correlation coefficient. The first three measures have been used previously in the context of document clustering [13,46] while the fourth has been used in other clustering applications [3].

1) Misclassification Rate: The misclassification (MC) rate, denoted by ν, is the proportion of documents that are classified incorrectly by the consensus clustering algorithm across all clusters based on the final cluster labels assigned by that algorithm. The MC rate can be calculated only if the true number of classes K is known and thus provides us with an overall measure of agreement for the clustering.

2) Normalized Mutual Information: The normalized mutual information (NMI) between the true class labels X and the assigned cluster labels Y is defined as

$$NMI = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}}$$

where $H(X)$ and $H(Y)$ represent the corresponding entropies [48]. For final cluster label assignments based on multiple runs, the estimate of NMI is given by

$$NMI = \frac{\sum_{i,j} n_{ij} \log\left(\frac{n \cdot n_{ij}}{n_i n_j}\right)}{\sqrt{\left(\sum_i n_i \log\frac{n_i}{n}\right)\left(\sum_j n_j \log\frac{n_j}{n}\right)}}$$

where n_i is the number of documents in class i, n_j is the number of documents in cluster j, n_{ij} is the number of documents in class i and cluster j and n is the total number of documents.

3) Adjusted Rand Index: The Adjusted Rand Index (ARI) is another commonly used measure to quantify the agreement between the true class labels X and the assigned cluster labels Y. It is the proportion of pairs of documents that are both in the same class and same cluster or that are both in a different class and different cluster, normalized to fall in the $[0, 1]$ range. For details on this measure, the interested reader is referred to [37] and references cited therein.

4) Cophenetic Correlation Coefficient: The cophenetic correlation coefficient ρ is defined as the correlation between $1 - \bar{C}$ (where $\bar{C}$ is the consensus matrix defined earlier) and the distance induced by HC using average linkage [3].

Unlike ν, NMI, ARI and ρ can be computed even if the true number of classes K is not known. However, for our purpose of evaluating clustering accuracy the true K is known. The range of each measure is $[0, 1]$ where the two extreme values correspond to random partitioning and perfect clustering, respectively. This enables us to compare these three measures by correlating each with ν across the range of γ based on the true K. We applied $N = 200$ runs of the algorithm for various choices of γ in the interval $(0, 2]$. For given K and each choice of γ, we compute each measure based on the final cluster assignments from the consensus clustering algorithm.

B. Sparseness

For each (k, γ) combination, we investigated the sparseness of the metaterms (columns of W) and metaterm frequency profiles (rows of H). The sparseness of a non-negative vector $\mathbf{x}$ of length n, denoted by $\psi(\mathbf{x})$, is given by [23]

$$\psi(\mathbf{x}) = \frac{\sqrt{n} - \frac{(\sum_i x_i)}{\sqrt{\sum_i x_i^2}}}{\sqrt{n} - 1}.$$

It is interesting to note that $\psi(\mathbf{x}) = 0$ if and only if all elements of $\mathbf{x}$ are equal and $\psi(\mathbf{x}) = 1$ if and only if $\mathbf{x}$ contains a single non-zero element.

C. Data Normalization

We consider two different normalization schemes for the term frequency matrix in our presentation of real examples and simulated data. These are term frequency normalization (*tf*) and term frequency-inverse document frequency normalization (*tfidf*). In the *tf* scheme, only the term frequencies are normalized to prevent any bias towards documents containing more terms. In *tfidf* normalization, the inverse document frequency (*idf*) which measures the general importance of a term is computed for each term. The *idf* for each term is then multiplied with the corresponding term frequencies to obtain *tfidf* normalized data. A detailed account of these methods can be found in [44] and [45]. Henceforth, we shall refer to these methods simply as *tf* and *tfidf*, respectively. For each dataset presented, we compare these two schemes based on clustering accuracy for each (k, γ) combination as well as their overall performance.

D. Algorithm Implementation

For any given value of the parameter γ in Renyi's divergence, the algorithm groups the samples into k clusters, where k is the pre-specified rank of the factorization. Our procedure requires us to evaluate various choices of γ for a fixed k, each based on N runs, by computing the corresponding misclassification rates. For any real large-scale data set, the implementation of the steps in this evaluation procedure for the combination of a given k and each γ can be computationally very intensive. However, the stochastic nature of the NMF algorithm enables each step in this procedure to be run independently and simultaneously. These steps can be repeated for each run of the algorithm and the information from independent runs combined via the consensus clustering algorithm. Thus the NMF algorithm lends itself easily to a parallel implementation that would greatly increase speed and efficiency. Recently, we discussed such a comprehensive parallel implementation of this algorithm on a Message-Passing Interface (MPI)/C++ platform [24] using high-performance computing (HPC) clusters [10]. We also created an integrated package with a graphical user interface that communicates between a Windows desktop and the HPC cluster using MPI compatible software on computer clusters. This implementation was utilized in all our computations.

VI. Real-life and Simulated Examples

We describe several real-life and simulated examples to illustrate the applicability of our methods as well as their performance. For this purpose, we consider the following choices of γ in the interval $(0, 2]$: $0.01, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2$, for various ranks k, each based on $N = 200$ runs. Note that $\gamma = 1$ represents KL divergence and $\gamma = 0.01$ approximates dual KL divergence described in (3.4). Other known measures considered here include the Bhattacharya distance $(\gamma = 0.5)$ and the Pearson chi-squared statistic $(\gamma = 2)$. We utilize two publicly available benchmark datasets for this purpose - the WebKB [25] and Reuters [46] data sets. In addition, we consider the Page Blocks data [16] to further illustrate our methodology to other document analysis problems. We present only the most relevant results from our analyses.

A. *WebKB Data*

The documents in the WebKB corpus are webpages collected from computer science departments of various universities by the World Wide Knowledge Base (WebKb) project and is available from [25]. It consists of 2803 documents split into four classes, namely, project, course, faculty and student. We pre-processed this dataset based on document and term frequencies alone. This resulted in 902 documents containing 1338 terms across the four classes.

Given four major classes of documents in this corpus, we considered a rank $k = 4$ factorization for the choices of γ listed earlier. Figure 1(a) displays the misclassification rate ν plotted as a function of γ for normalized data based on *tf* and *tfidf*. These normalization methods are seen to perform similarly overall where $\gamma = 0.75$ and 0.5 result in the lowest misclassification rate of 25.1% and 25.2%, respectively. This is corroborated by the relationship between ARI and γ as well as between NMI and γ, and is graphically presented in Figures 1 (b) & (c). The strong negative correlation between each of these measures and ν is evidenced in Figures 1 (e) & (f) where, for each normalization method, the most homogeneous cluster corresponds to the value of γ that results in the smallest misclassification rate. The relationship between ρ and γ is similar to that observed for ARI and NMI (Figures 1 (d)). However, the correlation between ρ and ν is not as strong as that observed for the other measures (Tables 3 and 4).

It is worth noting that $\gamma = 1$ (KL divergence) results in higher misclassification rates of 26.2% and 29.4% using *tf* and *tfidf*, respectively. Using similar measures of clustering accuracy and a different filtered version of this dataset, reference [13] demonstrated superior performance of

their NMF-based hybrid method. They used only the top 1000 terms in the corpus selected based on mutual information with class labels. On the other hand, our filtering scheme is completely blinded to the class labels and utilizes only the term and document frequencies. Nevertheless, our approach achieves a clustering accuracy of 75% using both *tf* ($\gamma = 0.75$) and *tfidf* ($\gamma = 0.5$) normalization (Figure 1(a)) and outperforms the 64.4% accuracy achieved by their hybrid method.

Furthermore, we investigated the sparseness of the four metaterms and meteterm frequency profiles for each choice of γ. For fixed γ, the mean sparseness of the metaterms (columns of W) was computed as the sparseness of each metaterm averaged across the four metaterms for each run and then averaged across the $N = 200$ runs. The mean sparseness of the metaterm frequency profiles (rows of H) was computed in a similar manner. For both normalization methods, sparseness of metaterms showed a monotonically decreasing trend with respect to γ (Figure 1(h)) while sparseness of metaterm frequency profiles showed an initial surge for small γ before declining for higher values of γ (Figure 1(g)). It is interesting to note that *tf* normalization resulted in uniformly sparser metaterms (across the range of γ) while *tfidf* normalization resulted in uniformly sparser metaterm frequency profiles.

B. Reuters Data

The Reuters data is one of the most widely used benchmark datasets in text mining. We utilize the pre-processed data consisting of the frequencies of 1969 terms from 276 different documents presented by [46,47]. These documents belong to a total of 20 different categories. For the purpose of illustrating our methodology, we created various subsets of this dataset where the known, true number K, of classes varied anywhere from 3 to 20. This allowed us to evaluate the performance of our method for various models, each determined by the true number of classes of documents. In each case, the appropriate rank of factorization was used.

Table 1: Subsets of Reuters Data

Subset	K	# of documents	# of terms
1	3	77	1969
2	4	68	1105
3	5	120	1527
4	6	139	1639
5	8	169	1706
6	10	195	1800
7	20	276	1969
8	4	99	1407
9	3	55	828

Table 1 presents a summary of the various subsets used in our analysis. Subsets 1 through 6 were created based on the K most frequently occurring classes of documents in the corpus where the corresponding K is specified in this table. Subset 7 represents the complete dataset. Subset 8 consists of the two most frequently occurring classes of documents and two less frequently occurring classes while subset 9 consists of the three most frequently occurring classes that follow the first four most frequently occurring classes. Each subset was first created by appropriate filtering and then normalized by *tf* and *tfidf* separately.

Figure 2 presents the misclassification rates for each subset plotted against γ for the two normalization methods. In most cases, *tf* performs at least as well as or better than *tfidf* in delineating the true classes. This includes the complete dataset represented by subset 7. In addition, the performance of *tf* appears to be fairly uniform across γ and better than *tfidf* for subsets 1 and 8. Interestingly for subsets 2 and 9, *tfidf* is the best performer even though *tf* appears to dominate *tfidf* for most of the range of γ. Perfect clustering ($\nu = 0$) is achieved for subset 1 using both *tf* ($\gamma = 0.5, 0.75$) and *tfidf* ($\gamma = 0.25$) methods; and for subset 8 using *tf* ($\gamma = 0.5$). In all subsets, there is at least one value of γ that outperforms $\gamma = 1$ (KL divergence). These observations are further corroborated by the relationship between ARI and γ as well as between NMI and γ as shown in Supplemental Figures S1 and S2, respectively. These two measures

display a similar pattern of change with respect to γ compared to that of ρ (Supplemental Figure S3).

Strong negative correlations between ARI and ν, and between NMI and ν are clearly seen for most subsets across both normalization methods (Tables 3 and 4). However, the correlations between ρ and ν are seen to be weaker overall. In particular, for subsets 3,4,6,8 and 9, a high value of ρ was associated with the largest misclassification rate for *tfidf* (Figure 2 and Supplemental Figure S3, panels (c),(d),(f),(g),(h)). In some cases (subsets 3 and 6), the highest ρ corresponds to the largest ν. This is corroborated by the poor (and sometimes positive) correlations for these cases (Table 4). Overall, ARI appears to have a stronger correlation with ν relative to other measures and *tf* normalization shows a stronger correlation with ν across all measures. Once again, a monotonically decreasing trend was observed in the sparseness of metaterms with respect to γ (Supplemental Figure S5) for both normalization methods. On the other hand, sparseness of metaterm frequency profiles showed an overall decreasing trend with increasing γ (Supplemental Figure S4). In most cases, *tf* normalization resulted in uniformly sparser metaterms while *tfidf* normalized data resulted in uniformly sparser metaterm frequency profiles across the range of γ.

The various subsets in this example allowed us to perform a sensitivity analysis with real data whereby we have assessed performance of our method for various true models. References [46,47] adopted a similar approach for evaluating their penalized NMF (PNMF) algorithm using this dataset. They considered ranks (subsets) ranging from $K = 2$ to 20 and assessed the clustering accuracy of their method for various choices of their penalty parameter λ. For more details, the interested reader is referred to their paper referenced above. It is not clear exactly how the subsets were chosen in their approach, nevertheless, it provides us with a basis for comparing the two methods for this dataset. Table 2 presents the clustering accuracies (1-ν expressed in %) for the two methods for various ranks. In each case, the best performing model determined by the choice(s) of γ or λ is listed. It is evident from these results that our approach performs better throughout the range of K considered. There is also a notable improvement in performance for smaller ranks where our method achieved near-perfect or perfect clustering. Interestingly, both methods show an improvement in clustering accuracy at $K = 10$ relative to $K = 8$ before declining for $K = 20$. Furthermore, reference [13] used the ten most frequently occurring categories in this dataset to demonstrate superior performance of their

NMF-based hybrid algorithm. Once again, they applied an informative filter that utilized the top 1000 terms based on mutual information with class labels. However, our approach performed significantly better (71% and 70% clustering accuracy using *tf* and *tfidf*, respectively) in clustering the documents over their hybrid approach (52.1% clustering accuracy).

Table 2: Comparison of Results: Reuters Data

K	$PNMF(\lambda)$	*tf*(γ)	*tfidf*(γ)
4	78 (0.001)	100 (0.5)	97 (0.75)
6	73 (0.001)	79 (1.75)	68 (0.75)
8	57 (0.1)	67 (0.5)	62 (0.75,1.25)
10	67 (0.01)	71 (1.0)	70 (1.25)
20	57 (0.001)	58 (1.5)	58 (1.5)

We developed exploratory tools for visualizing the clustering using the metaterms and metaterm frequency profiles. For the purpose of illustration, we consider subset 8 consisting of four classes of documents. Perfect clustering $(\nu = 0)$ was achieved for this data for the case $\gamma = 0.5$ based on *tf* normalization (see Figure 2(h)). The sparseness of the metaterms is best illustrated by a box-plot of its coefficients, shown in Supplemental Figure S6 for this subset. The coefficient of a given term in each metaterm quantifies the influence of that term in the corresponding metaterm frequency profile of documents in a corpus. The compressed distribution of the metaterms highlights terms with relatively high coefficients. We identified a core set of 50 terms that appear in the top 5% of each of the four metaterms; as well as a unionized set of 137 terms that appear in the top 5% of at least one metaterm. This overlapping nature of the metaterms illustrates the potential role played by the frequency of occurrence of a single term in multiple classes of documents.

A plot of these metaterm frequency profiles can help illustrate the role played by each metaterm. Figure 3 presents the mean frequency profiles (across the $N = 200$ runs) for the four metaterms, i.e., a plot of the rows of H averaged across the $N = 200$ runs. The numbers 1-4 in this figure represent the four classes. It is evident from these profiles that each metaterm aids

in delineating exactly one class from the rest. For example, the first metaterm profile separates class 4 from the rest and so on. Even though the class labels of documents are known in this dataset and have been plotted in order, the separation provided by each metaterm profile is clear. A visual summary of the performance of clustering can thus be obtained using such a plot.

C. Page Blocks Data

This dataset was described by [16] and represents a unique example in document analysis. Here, we are interested in classifying all the blocks of the page layout of a document that have been detected by a segmentation process. This is an important step in document analysis that is necessary for separating text from non-text areas. The original dataset consists of 5473 blocks from 54 distinct documents. Each block represents an observation and there are five classes of blocks, namely, text, horizontal line, picture, vertical line and graphic. The following variables are measured for each block - number of black pixels per unit area, mean number of white-black transitions, total number of black pixels, number of white-black transitions in the original bitmap of the block, height, length, area and eccentricity (ratio of length to height). In addition, the dataset also contains the number of black pixels per unit area and the total number of black pixels obtained after the application of a smoothing algorithm. This dataset is available from [26]. For more details on this dataset, the interested reader is referred to [16] and [34].

We reduced the dimensionality of this dataset by removing blocks with a relatively small number of black pixels per unit area *and* mean number of white-black transitions. This resulted in 1407 blocks across the five classes. Also since some variables have been normalized with respect to other variables in this dataset, no further normalization (i.e., tf or $tfidf$) was deemed necessary for this data. A rank $k = 5$ factorization was applied to this dataset for each γ under consideration. $\gamma = 0.25$ resulted in the most homogeneous grouping of blocks based on the measured variables. This is indicated by the observed trend in ARI and NMI as γ increases (Figures 4 (b) & (c)) and by the smallest misclassification rate, ν, of 29.6% achieved among all choices of γ (Figure 4(a)). On the other hand, ρ exhibits an inconsistent change as γ increases, peaking at $\gamma = 1$ (corresponding to $\nu = 41.8\%$) (Figure 4(d)). In particular, it is worth noting that $\gamma = 0.25$ outperforms all four known metrics - KL divergence $(\gamma = 1, \nu = 41.8\%)$, approximation to dual KL divergence $(\gamma = 0.01, \nu = 44.3\%)$, Bhattacharya distance $(\gamma = 0.5, \nu = 36.7\%)$ and the Pearson chi-squared statistic $(\gamma = 2, \nu = 46\%)$ - by a wide margin. These results emphasize

the need to incorporate different choices of γ in the factorization, beyond the commonly known metrics.

Table 3: Correlation between measures of clustering accuracy: *tf*

Dataset	ARI vs. ν	NMI vs. ν	ρ vs. ν
WebKB	-1.00	-0.98	-0.81
Reuters 1	-1.00	-0.98	-0.16
Reuters 2	0.24	-0.38	-0.77
Reuters 3	-0.82	0.05	0.36
Reuters 4	-0.96	-0.41	-0.66
Reuters 5	-0.87	-0.96	-0.33
Reuters 6	-0.98	-0.97	-0.80
Reuters 7	-0.98	-0.96	-0.88
Reuters 8	-1.00	-0.99	0.33
Reuters 9	-0.63	-0.43	-0.15
Page Blocks	-0.85	-0.19	0.03
Simulated 1	0.10	0.34	0.43
Simulated 2	-0.97	-0.95	0.53
Simulated 3	-0.99	-1.00	-0.47

D. Simulating Nested Classes

We further investigate the performance of the NMF algorithm via extensive simulations involving a correlated structure. In particular, we illustrate its ability to recover documents into the true underlying classes when there exists a sub-structure (or a dependent structure) between different classes. This is more realistic in real-life data especially when the number of classes exceeds two, and there is a hierarchical or nested structure of the classes. To this end, we construct three examples involving simulated frequencies of $p = 1000$ terms for each of $n = 60$ documents. We first describe their construction followed by their analyses based on our methods.

Table 4: Correlation between measures of clustering accuracy: *tfidf*

Dataset	ARI vs. ν	NMI vs. ν	ρ vs. ν
WebKB	-0.99	-0.97	-0.77
Reuters 1	-0.97	-0.95	-0.70
Reuters 2	-0.49	-0.55	-0.50
Reuters 3	-0.97	-0.93	0.10
Reuters 4	-0.85	-0.84	0.25
Reuters 5	-0.74	-0.96	-0.66
Reuters 6	-0.89	-0.94	-0.12
Reuters 7	-0.94	-0.93	-0.83
Reuters 8	-0.91	-0.91	-0.54
Reuters 9	-0.70	-0.45	-0.16
Page Blocks	-0.85	-0.19	0.03
Simulated 1	-0.25	0.64	0.65
Simulated 2	0.63	0.97	0.37
Simulated 3	-0.25	0.97	0.41

a) Example 1: We generated the term-document frequencies as follows: Let documents $1-20$, $21-40$ and $41-60$ denote classes A, B and C respectively. For the first 50 terms, frequencies for documents in classes A, B and C were generated from a Poisson distribution with means 10, 1 and 1 respectively. For terms $51-100$, frequencies for documents in class B were generated as $Y \sim \min(X_1, X_2)$ where $X_1 \sim Poisson(mean = \lambda_1)$ and $X_2 \sim Poisson(mean = \lambda_2)$; and frequencies for documents in class C were generated as $Z \sim \max(X_3, X_2)$ where $X_3 \sim Poisson(mean = \lambda_3)$. For terms $51-100$, documents in classes B and C have a dependent structure while for terms $1-50$, documents in class A are independent of those in classes B and C. For the remainder of the terms, all documents are generated from the same background distribution as before, i.e., Poisson with unit mean. We considered various combinations of λ_1, λ_2 and λ_3 in the following ranges: $\lambda_1 \in [50, 70], \lambda_2 \in [70, 100]$ and $\lambda_3 \in [20, 50]$.

b) Example 2: We generated toy data based on the same setup as Example 1 above except for the following: For terms $51-100$, frequencies for documents in class B were generated as $Y \sim min(X_1, X_2)$ where $X_1 \sim Poisson(mean = \lambda_1)$ and $X_2 \sim Poisson(mean = \lambda_2)$; and frequencies for documents in class C were generated from a Poisson distribution with mean λ_3. As in Example 2, documents in classes B and C have a dependent structure for terms $51-100$ while for terms $1-50$, documents in class A are independent of those in classes B and C. For the remainder of the terms, all documents are generated from a Poisson distribution with unit mean. We set $\lambda_1 = 10$ and considered various choices of λ_2 in the range $(20, 40]$.

c) Example 3: For this example, we generated toy data based on the same setup as Example 2 above except for the following: For the first 50 terms, frequencies for documents in classes A, B and C were generated from a Poisson distribution with means 20, 1 and 1 respectively. Document frequencies for terms 51-100 in classes B and C, as well as the remainder of the terms in all three classes were generated as described in Example 2. In this set-up, documents in classes B and C have a dependent structure for terms $51-100$ while documents in classes A and B have a dependent structure among the first 100 terms. In this structure, class B is dependent on both classes A and C. We set $\lambda_1 = 20$ and considered various choices of λ_2 in the range $[25, 40]$.

The rationale behind this scheme is to generate data with a dependent and/or a hierarchical structure. In the first two examples, there are two major classes where one class has two sub-classes while the third example represents a unique dependent structure between all three classes. Each dataset was normalized using *tf* and *tfidf* and then a rank $k = 3$ factorization was applied using our method.

The overall performance of our method on the simulated datasets paralleled that on the real-life data presented. The results were also insensitive to the choice of the Poisson mean parameters in each example. For the sake of brevity, we present results only for the case $\lambda_1 = 80, \lambda_2 = 50$ and $\lambda_3 = 20$ in Example 1; and for the case $\lambda_1 = 20, \lambda_2 = 25$ in Examples 2 and 3. Figure 5 presents the misclassification rates for each example plotted against γ for the two normalization methods. *tf* not only outperforms *tfidf* in delineating the true classes, its performance is also uniformly better than that of *tfidf* throughout the range of γ. There is also at least one value of γ that outperforms KL divergence ($\gamma = 1$) in each example for both *tf* and *tfidf*. In Examples 1 and 2, there is a significant improvement in clustering using *tf* for $\gamma = 0.5$ ($\nu = 25\%$) and

$\gamma = 1.5$ ($\nu = 3.33\%$), respectively, compared to $\gamma = 1$ ($\nu = 31.67\%$ and 26.67% respectively). The difference in performance between *tf* and *tfidf* is particularly striking in Example 3.

ARI, NMI and ρ displayed relationships with γ similar to those observed for real-life data (data not shown). NMI was poorly correlated with ν for *tfidf* while ρ was poorly correlated with ν for both normalization methods in all three examples above (Tables 3 and 4). We also observed a decreasing trend in the sparseness of both the metaterm frequency profiles and metaterms with respect to γ (Supplemental Figures S7 & S8) for both normalization methods. Once again, *tf* normalization resulted in sparser metaterms overall while *tfidf* normalized data resulted in almost uniformly sparser metaterm frequency profiles across the range of γ. We also considered a rank $k = 2$ factorization in Examples 1 and 2, and observed that the performance of our method was uniform across the range of γ (data not shown). This is evidently due to the relative homogeneity of individual classes caused by the data generating mechanism.

Next, we investigated the sensitivity of our method in delineating similar clusters. In our simulation studies, the similarity between any two clusters can be simply determined by the data generating mechnism. To this end, we utilized Example 2 and varied the Poisson mean parameter λ_2 that determines the degree of closeness between classes B and C. As λ_2 was decreased from 40 to about 20, i.e., as classes B and C became more and more similar, a gradual increase was observed in the misclassification rate across γ (see Figure 6 and Figure 5(b) for $\lambda_2 = 25$). Note that as $\lambda_2 \to 20$, classes B and C merge into a single, larger class. *tf* performs uniformly better than *tfidf* not only throughtout the range of γ but also that of λ_2. Moreover, the performance of *tfidf* appears to have reached saturation at $\lambda_2 = 30$ and does not show any further improvement as λ_2 is decreased. It is natural that *tf* performs very well for $\lambda_2 \geq 30$, however, the utility of our method lies in its ability to capture subtle differences between classes B and C (i.e., as λ_2 approaches the limiting value of 20). When $\lambda_2 = 25$, $\gamma = 1.5$ is the best-performing model with a misclassification rate of 3.33% while $\gamma = 1$ (KL divergence) is the worst-performing model with a misclassification rate of 26.67%. For $\lambda_2 = 22$, $\gamma = 0.25$ is the best performer with a misclassification rate of 16.67% compared to 28.33% for KL divergence. This phenomenon was also observed in other examples in our simulation studies as the parameter values were varied (data not shown) and it emphasizes the need for a broader approach.

VII. Summary and Discussion

In summary, we have described a unified algorithm for NMF based on Renyi's divergence and proved convergence of our algorithm using an auxiliary function analogous to that used for proving convergence of the EM algorithm. This approach provides a unique and generalized framework for NMF and includes well-known distance measures as special cases. Furthermore, we generalized the equivalence between NMF and PLSI using a Box-Cox transformation in the multinomial likelihood for PLSI. This generalization embeds PLSI within the larger framework of the λ-log-likelihood. Last but not least, we demonstrated the applicability of our methods using simulated as well as real-life document clustering data.

One of the objectives of this paper has been to demonstrate the need for a generalized metric for modeling high-dimensional data in the context of text mining and document clustering. The generalized metric presented here retains the distributional assumption on the data while providing modeling flexibility via the choice of the parameter γ. In that regard, one could arguably view our approach from the perspective of penalized likelihood where the choice of γ in Renyi's divergence determines the joint penalty on the metaterms and metaterm frequency profiles, or alternatively, on the reconstructed matrix WH. Furthermore, the application of consensus clustering to select γ is analogous to the use of cross-validation for choosing the penalty parameter in penalized likelihood methods.

Our real-life examples and simulation studies suggest an underlying effect due to the distribution of the term frequencies (across documents) on the performance of the clustering algorithm. This is determined by the choice of γ. Perfect clustering is indeed achievable with the appropriate choice of γ for some datasets, as demonstrated in our examples. The approach emphasizes the need for a data-driven choice of γ and, hence, of the divergence measure itself used in the decomposition. In practice, we recommend the use of several values of γ for evaluating the clustering accuracy for a given factorization rank k. We focused mainly on values of γ in $(0, 2]$ in this paper for illustrative purposes, however, other cases might be of interest and can be implemented easily. Our parallel implementation has distinct advantages in terms of computational speed and allows one to simultaneously evaluate several factorization ranks.

An important observation from the analytical results presented is that the best performing model is not necessarily the sparsest, either in terms of the metaterms or the metaterm frequency

profiles. The simulations highlight the ability of our approach to delineate classes based on subtle differences between them. The overall performance of *tf* normalization has been superior to that of *tfidf*. Even though ARI, NMI and ρ can be used as measures of cluster homogeneity even when the true number of clusters is unknown, our results demonstrated that both ARI and NMI were better measures of clustering accuracy than ρ. The problems associated with ρ have been well documented in the literature [18,20]. Moreover, ρ typically has too narrow a range to be useful in many applications and, unlike ARI and NMI, can only be used with consensus clustering in conjunction with hierarchical clustering. In particular, for a real dataset with unknown true number of classes, we recommend the use of ARI or NMI on *tf* normalized data.

While Renyi's divergence is directly applicable for modeling data generated from the Poisson distribution, it has also been successfully used for modeling large-scale biological data such as those from microarray studies [8,12]. It has been shown to closely approximate data from skewed distributions in such studies. There is also evidence suggesting that a heavy-tailed continuous distribution such as the gamma distribution closely approximates data from many microarray studies [12]. A useful resource on this topic is [27]. Applications of NMF involving special cases such as Kullback-Leibler divergence and Euclidean distance in the domain of computational biology are abundant in the literature. An extensive and fairly recent list of such applications is presented in [12]. Thus the approach presented here provides the generalizability and flexibility in modeling such large-scale biological data as well, and further broadens the usefulness and applicability of our method.

APPENDIX I
PROOF OF THOEREM 1

First, we prove this result for $0 < \gamma < 1$. Then we show how similar arguments can be used to prove the result for $\gamma > 1$ and for $\gamma < 0$.

We will make use of an auxiliary function similar to the one used in the Expectation-Maximization (EM) algorithm [7,32]. Note that for h real, $G(h, h')$ is an auxiliary function for $F(h)$ if $G(h, h') \geq F(h)$ and $G(h, h) = F(h)$ where G and F are scalar valued functions. Also, if G is an auxiliary function, then F is non-increasing under the update $h^{t+1} = \arg\min_h G(h, h^t)$.

We define

$$F(H_{ak}) = \gamma \sum_i V_{ik} + (1-\gamma) \sum_{i,a} W_{ia} H_{ak} - \sum_i V_{ik}^{\gamma} \left[\sum_a W_{ia} H_{ak} \right]^{1-\gamma},$$

where H_{ak} denotes the ak^{th} entry of H. Then the auxiliary function for $F(H_{ak})$ is

$$G(H_{ak}, H_{ak}^t) = \gamma \sum_i V_{ik} + (1-\gamma) \sum_{i,a} W_{ia} H_{ak} - \sum_{i,a} V_{ik}^{\gamma} \frac{W_{ia} H_{ak}^t}{\sum_b W_{ib} H_{bk}^t}$$
$$\left[(W_{ia} H_{ak})^{1-\gamma} \left(\frac{W_{ia} H_{ak}^t}{\sum_b W_{ib} H_{bk}^t} \right)^{\gamma-1} \right]. \tag{1.15}$$

It is straightforward to show that $G(H_{ak}, H_{ak}) = F(H_{ak})$. To show that $G(H_{ak}, H_{ak}^t) \geq F(H_{ak})$, we use the convexity of $-x^{1-\gamma}$ and the fact that for any convex function $f, f\left(\sum_{i=1}^n r_i x_i\right) \leq \sum_{i=1}^n r_i f(x_i)$ for rational nonnegative numbers $r_1, \cdots, r_n$ such that $\sum_{i=1}^n r_i = 1$. We then obtain

$$-\left(\sum_a W_{ia} H_{ak} \right)^{1-\gamma} \leq -\sum_a \gamma_a \left(\frac{W_{ia} H_{ak}}{\gamma_a} \right)^{1-\gamma} =$$
$$-\sum_a (W_{ia} h_a)^{1-\gamma} \left(\frac{W_{ia} H_{ak}^t}{\sum_b W_{ib} H_{bk}^t} \right) \frac{(\sum W_{ib} H_{bk}^t)^{1-\gamma}}{(W_{ia} H_{ak}^t)^{1-\gamma}}$$

where $\gamma_a = \dfrac{W_{ia} H_{ak}^t}{\sum_b W_{ib} H_{bk}^t}$. From this inequality it follows that $F(H_{ak}) \leq G(H_{ak}, H_{ak}^t)$.

Now, $\dfrac{dG(H_{ak}, H_{ak}^t)}{dH_{ak}} = (1-\gamma) \sum_i W_{ia} -$

$$\sum_i (1-\gamma) V_i^{\gamma} \left(W_{ia}^{1-\gamma} \right) \frac{W_{ia} H_{ak}^t}{\sum_b W_{ib} H_{bk}^t} \left(\frac{W_{ia} H_{ak}^t}{\sum_b W_{ib} H_{bk}^t} \right)^{\gamma-1} H_{ak}^{-\gamma} = 0$$

Thus, the update rule for H takes the form

$$H_{ak}^{t+1} = \left(\frac{\sum_i V_{ik}^{\gamma} W_{ia}^{1-\gamma} (W_{ia} H_{ak}^t)^{\gamma}}{\left(\sum_b W_{ib} H_{bk}^t \right)^{\gamma} \sum_i W_{ia}} \right)^{1/\gamma}$$
$$= H_{ak}^t \left(\frac{\sum_i \left(\frac{V_{ik}}{\sum_b W_{ib} H_{bk}^t} \right)^{\gamma} W_{ia}}{\sum_i W_{ia}} \right)^{1/\gamma}.$$

Similarly, we define

$$F(W_{ia}) = \gamma \sum_i V_{ik} + (1-\gamma) \sum_{i,a} W_{ia} H_{ak} - \sum_i V_{ik}^{\gamma} \left[\sum_a W_{ia} H_{ak} \right]^{1-\gamma},$$

where W_{ia} denotes the ia^{th} entry of W. Then the auxiliary function for $F(W_{ia})$ is

$$G(W_{ia}, W_{ia}^t) = \gamma \sum_i V_{ik} + (1-\gamma) \sum_{i,a} W_{ia} H_{ak} - \sum_{i,a} V_{ik}^{\gamma} \frac{W_{ia}^t H_{ak}}{\sum_b W_{ib}^t H_{bk}} \left[(W_{ia} H_{ak})^{1-\gamma} \left(\frac{W_{ia}^t H_{ak}}{\sum_b W_{ib}^t H_{bk}} \right)^{\gamma-1} \right]. \tag{1.16}$$

It is straightforward to show that $G(W_{ia}, W_{ia}) = F(W_{ia})$. To show that $G(W_{ia}, W_{ia}^t) \geq F(W_{ia})$, we use the convexity of $-x^{1-\gamma}$ and the fact that for any convex function $f, f\left(\sum_{i=1}^n r_i x_i\right) \leq \sum_{i=1}^n r_i f(x_i)$ for rational nonnegative numbers $r_1, \cdots, r_n$ such that $\sum_{i=1}^n r_i = 1$. We then obtain

$$-\left(\sum_a W_{ia} H_{ak} \right)^{1-\gamma} \leq -\sum_a \gamma_a \left(\frac{W_{ia} H_{ak}}{\gamma_a} \right)^{1-\gamma} =$$

$$-\sum_a (W_{ia} h_a)^{1-\gamma} \left(\frac{W_{ia}^t H_{ak}}{\sum_b W_{ib}^t H_{bk}} \right) \frac{(\sum W_{ib}^t H_{bk})^{1-\gamma}}{(W_{ia}^t H_{ak})^{1-\gamma}}$$

where $\gamma_a = \dfrac{W_{ia}^t H_{ak}}{\sum_b W_{ib}^t H_{bk}}$. From this inequality it follows that $F(W_{ia}) \leq G(W_{ia}, W_{ia}^t)$.

Now, $\dfrac{dG(W_{ia}, W_{ia}^t)}{dW_{ia}} = (1-\gamma) \sum_a H_{ak} -$

$$\sum_i (1-\gamma) V_i^{\gamma} \left(H_{ak}^{1-\gamma} \right) \frac{W_{ia}^t H_{ak}}{\sum_b W_{ib}^t H_{bk}} \left(\frac{W_{ia}^t H_{ak}}{\sum_b W_{ib}^t H_{bk}} \right)^{\gamma-1} W_{ia}^{-\gamma} = 0$$

Thus, the update rule for W takes the form

$$W_{ia}^{t+1} = \left(\frac{\sum_k V_{ik}^{\gamma} H_{ak}^{1-\gamma} (W_{ia}^t H_{ak})^{\gamma}}{\left(\sum_b W_{ib}^t H_{bk} \right)^{\gamma} \sum_i H_{ak}} \right)^{1/\gamma}$$

$$= W_{ia}^t \left(\frac{\sum_k \left(\frac{V_{ik}}{\sum_b W_{ib}^t H_{bk}} \right)^{\gamma} H_{ak}}{\sum_k H_{ak}} \right)^{1/\gamma}.$$

This completes the proof for the case $0 < \gamma < 1$.

For $\gamma > 1$, using (3.5) we define

$$F(H_{ak}) = -\gamma \sum_i V_{ik} - (1-\gamma) \sum_{i,a} W_{ia} H_{ak} + \sum_i V_{ik}^{\gamma} \left[\sum_a W_{ia} H_{ak} \right]^{1-\gamma}$$

and the auxiliary function for $F(H_{ak})$ as

$$G(H_{ak}, H_{ak}^t) = -\gamma \sum_i V_{ik} - (1-\gamma) \sum_{i,a} W_{ia} H_{ak} + \sum_{i,a} V_{ik}^{\gamma} \frac{W_{ia} H_{ak}^t}{\sum_b W_{ib} H_{bk}^t}$$
$$\left[(W_{ia} H_{ak})^{1-\gamma} \left(\frac{W_{ia} H_{ak}^t}{\sum_b W_{ib} H_{bk}^t} \right)^{\gamma - 1} \right]. \tag{1.17}$$

It is easy to see that $G(H_{ak}, H_{ak}) = F(H_{ak})$. By using the convexity of $x^{1-\gamma}$ for $\gamma > 1$, we can show that $F(H_{ak}) \leq G(H_{ak}, H_{ak}^t)$ and proceed to obtain the update rules for H and W as described above. The update rules for this case are exactly as those specified for the case $0 < \gamma < 1$.

Finally, the proof for the case $\gamma < 0$ is obtained by using (3.6) and defining

$$F(H_{ak}) = -(1-\gamma) \sum_i V_{ik} - \gamma \sum_{i,a} W_{ia} H_{ak} + \sum_i V_{ik}^{1-\gamma} \left[\sum_a W_{ia} H_{ak} \right]^{\gamma}$$

and the auxiliary function for $F(H_{ak})$ to be

$$G(H_{ak}, H_{ak}^t) = -(1-\gamma) \sum_i V_{ik} - \gamma \sum_{i,a} W_{ia} H_{ak} + \sum_{i,a} V_{ik}^{1-\gamma} \frac{W_{ia} H_{ak}^t}{\sum_b W_{ib} H_{bk}^t}$$
$$\left[(W_{ia} H_{ak})^{\gamma} \left(\frac{W_{ia} H_{ak}^t}{\sum_b W_{ib} H_{bk}^t} \right)^{-\gamma} \right]. \tag{1.18}$$

Again, it is easy to verify that $G(H_{ak}, H_{ak}) = F(H_{ak})$. Using the convexity of x^{γ} for $\gamma < 0$, we can show that $F(H_{ak}) \leq G(H_{ak}, H_{ak}^t)$ and proceed to obtain the update rules for H and W as shown above.

ACKNOWLEDGEMENT

Research of the the first author was supported in part by NIH Grant PA CA 06297 and an appropriation from the Commonwealth of Pennsylvania. The authors thank Prof. Michael Berry at the University of Tennessee, Knoxville for kindly provding the Reuters dataset. The authors also wish to acknowledge the assistance of Joseph Anlage of the High-Performance Computing Facility at Fox Chase Cancer Center.

REFERENCES

[1] AGRESTI, A. (1990). Categorical Data Analysis. John Wiley, New York.

[2] BOX, G. E. P., COX, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society, Series B 26 (2): 211252.

[3] BRUNET, J-P., TAMAYO, P., GOLUB, T., and MESIROV, J. (2004). Metagenes and molecular pattern discovery using nonnegative matrix factorization, *Proceedings of the National Academy of Sciences*, 101, 4164-4169.

[4] CHAGOYEN M, CARMONA-SAEZ P, SHATKAY H, CARAZO JM, PASCUAL-MONTANO A (2006) Discovering semantic features in the literature: a foundation for building functional associations, BMC Bioinformatics, 7:41.

[5] CRESSIE, N. and READ, T. R. C. (1984). Multinomial goodness-of-fit tests. Journal of the Royal Statistical Society, Series B, 46:440-464.

[6] CRESSIE, N., PARDO, L. and PARDO, M. (2003). Size and power considerations for testing log-linear models using ϕ-divergence test statistics, Statistica Sinica, 13, 555-570.

[7] DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, 39, 1-38.

[8] DEVARAJAN, K. and EBRAHIMI, N. (2005). Molecular pattern discovery using nonnegative matrix factorization based on Renyi's information measure. In: Proceedings of the XII SCMA International Conference; 24 December 2005; Auburn, Alabama (http://atlas-conferences.com/c/a/q/t/98.htm).

[9] DEVARAJAN, K. (2006). Nonnegative matrix factorization - A new paradigm for large-scale biological data analysis, Proceedings of the Joint Statistical Meetings, Seattle, Washington.

[10] DEVARAJAN, K. and WANG, G. (2007). Parallel implementation of non-negative matrix algorithms using high-performance computing cluster. In: Proceedings of the 39th Symposium on the Interface: Computing Science and Statistics. Theme: Systems Biology. 39th Symposium on the Interface: Computing Science and Statistics; 2326 May 2007; Temple University, Philadelphia, Pennsylvania (http://sbm.temple.edu/ interface07/documents/Interface07InvitedProgram_Revised.pdf).

[11] DEVARAJAN, K. (2008). Nonnegative matrix factorization - An analytical and interpretive tool in computational biology, PLoS Computational Biology, 4(7): e1000029. doi:10.1371/journal.pcbi.1000029.

[12] DEVARAJAN K. and EBRAHIMI, N. (2008). Class discovery via nonnegative matrix factorization. American Journal of Management and Mathematical Sciences, 28(3&4): 457-467.

[13] DING, C., LI, T., PENG, W. (2008). On the equivalence between nonnegative matrix factorization and probabilistic latent semantic indexing, Computational Statistics and Data Analysis, 52: 39133927.

[14] DING, C., LI, T., PENG, W. (2006). Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence, chi-square statistic, and a hybrid method, In: Proc. of National Conf. on Artificial Intelligence, AAAI-06, 2006.

[15] EBRAHIMI, N. and SOOFI, E. (2004). Information functions for Reliability. In Soyer, R., Mazzuchi, T.A. and Singpurwalla, N.D. (eds), *Mathematical Reliability, An Expository Perspective*. Kluwer's International, 127-159.

[16] ESPOSITO F., MALERBA D. and SEMERARO G., Multistrategy Learning for Document Recognition, Applied Artificial Intelligence, 8, pp. 33-84, 1994.

[17] FREEMAN, M.F. and TUKEY, J.W. (1950). Transformations related to the angular and the square root, Annals of Mathematical Statistics, 21, 607-611.

[18] HASTIE, T., TIBSHIRANI, R., and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*, New York: Springer-Verlag.

[19] HOFFMAN, T. (1999). Probabilistic Latent Semantic Analysis, Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence.

[20] HOLGERSSON, M. (1978). The limited value of cophenetic correlation as a clustering criterion, Pattern Recognition, 10(4):287-295.

[21] HOYER PO (2002) Nonnegative sparse coding. Neural Networks for Signal Processing XII 557565. IEEE Workshop on Neural Networks for Signal Processing; 46 September 2002; Martigny, Switzerland. 24.

[22] HOYER PO (2003) Modeling receptive fields with nonnegative sparse coding. Neurocomputing 5254: 547552. 25.

[23] HOYER PO (2004) Nonnegative matrix factorization with sparseness constraints. Journal of Machine Learning Research, 5: 14571469.

[24] http://www-unix.mcs.anl.gov/mpi/mpich2/

[25] http://web.ist.utl.pt/∼acardoso/datasets/

[26] http://archive.ics.uci.edu/ml/datasets/Page+Blocks+Classification

[27] http://discover.nci.nih.gov/microarrayAnalysis/

[28] KIM, H. and PARK, H. (2006). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares. In: Proceedings of the IASTED International Conference on Computational and Systems Biology 95100. IASTED International Conference on Computational and Systems Biology; 1314 November 2005; Dallas, Texas.

[29] KULLBACK, S. (1959). *Information Theory and Statistics*, New York: Wiley.

[30] KULLBACK, S. and LEIBLER, R.A. (1951). On information and sufficiency, *The Annals of Mathematical Statistics*, 22, 79-86.

[31] LEE, D.D., and SEUNG, S.H. (1999). Learning the parts of objects by nonnegative matrix factorization, *Nature*, 401, 788-791.

[32] LEE, D.D., and SEUNG, S.H. (2001). Algorithms for nonnegative matrix factorization, *Advances in Neural Information Processing Systems*, 13, 556-562.

[33] LIN, C-J. (2007). Projected gradient methods for non-negative matrix factorization. Neural Computation, 19:2756-2779.

[34] MALERBA, D., ESPOSITO, F. and SEMERARO, G., A Further Comparison of Simplification Methods for Decision-Tree Induction, In D. Fisher and H. Lenz (Eds.), "Learning from Data: Artificial Intelligence and Statistics V", Lecture Notes in Statistics, Springer Verlag, Berlin, 1995.

[35] MATSUYAMA, Y. (2003). The α-EM Algorithm: Surrogate likelihood maximization using α-logarithmic information measures, IEEE Transactions on Information Theory, 49(3): 692-706.

[36] MATUSITA, K. (1954). On estimation by the minimum distance method, Annals of the Institute of Statistical Mathematics, 5, 59-65.

[37] MONTI, S., TAMAYO, P., GOLUB, T.R., and MESIROV, J.P. (2003). Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data, *Machine Learning Journal*, 52, 91-118.

[38] NEYMAN, J. (1949). Contributions to the theory of the χ^2 test, *Proceedings of the First Berkeley Smposium on Mathematical Statistics and Probability*, Berkeley, University of California Press.

[39] PAUCA, P., SHAHNAZ, F., BERRY, M., and PLEMMONS, R. (2004). Text mining using nonnegative matrix factorization, *Proceedings of the SIAM International Conference on Data Mining, Orlando, April 2004*.

[40] PASCUAL-MONTANO, P., CARAZO, J.M., KOCHI, K., LEHMANN, D., PASCUAL-MARQUI, R. (2006) Nonsmooth nonnegative matrix factorization, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(3):403-415.

[41] R DEVELOPMENT CORE TEAM (2005). R: A language and environment for statistical computing, *R Foundation for Statistical Computing, Vienna, Austria*. ISBN 3-900051-07-0, URL http://www.R-project.org.

[42] READ, T. R. C. and CRESSIE, N. A. C. (1988). Goodness-of-fit Statistics for Discrete Multivariate Data. Springer-Verlag, New York.

[43] RENYI, A. (1970). *Probability Theory*, Amsterdam: North Holland.

[44] SALTON, G. and BUCKLEY, C. (1988). Term-weighting approaches in automatic text retrieval, Information Processing & Management 24 (5): 513523.

[45] SALTON, G. and MCGILL, M.J. (1983). Introduction to modern information retrieval, McGraw-Hill. ISBN 0070544840.

[46] SHAHNAZ, F. and BERRY, M. (2004). Document clustering using nonnegative matrix factorization, *Technical Report 2004-7, Department of Mathematics, Wake Forest University, North Carolina.*

[47] SHAHNAZ, F., BERRY, M., PAUCA, V.P. and R.J. Plemmons (2006). Document clustering using nonnegative matrix factorization, Information Processing and Management: An International Journal, 42(2): 373-386.

[48] STREHL, A. and GHOSH, J. (2002). Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions, Journal of Machine Learning Research, 3:583-617.

[49] WANG, G., KOSSENKOV A.V. and OCHS, M.F. (2005). LS-NMF: A modified non-negative matrix factorization algorithm utilizing uncertainty estimates. BMC Bioinformatics 7: 175.

Figure 1: WebKB Data

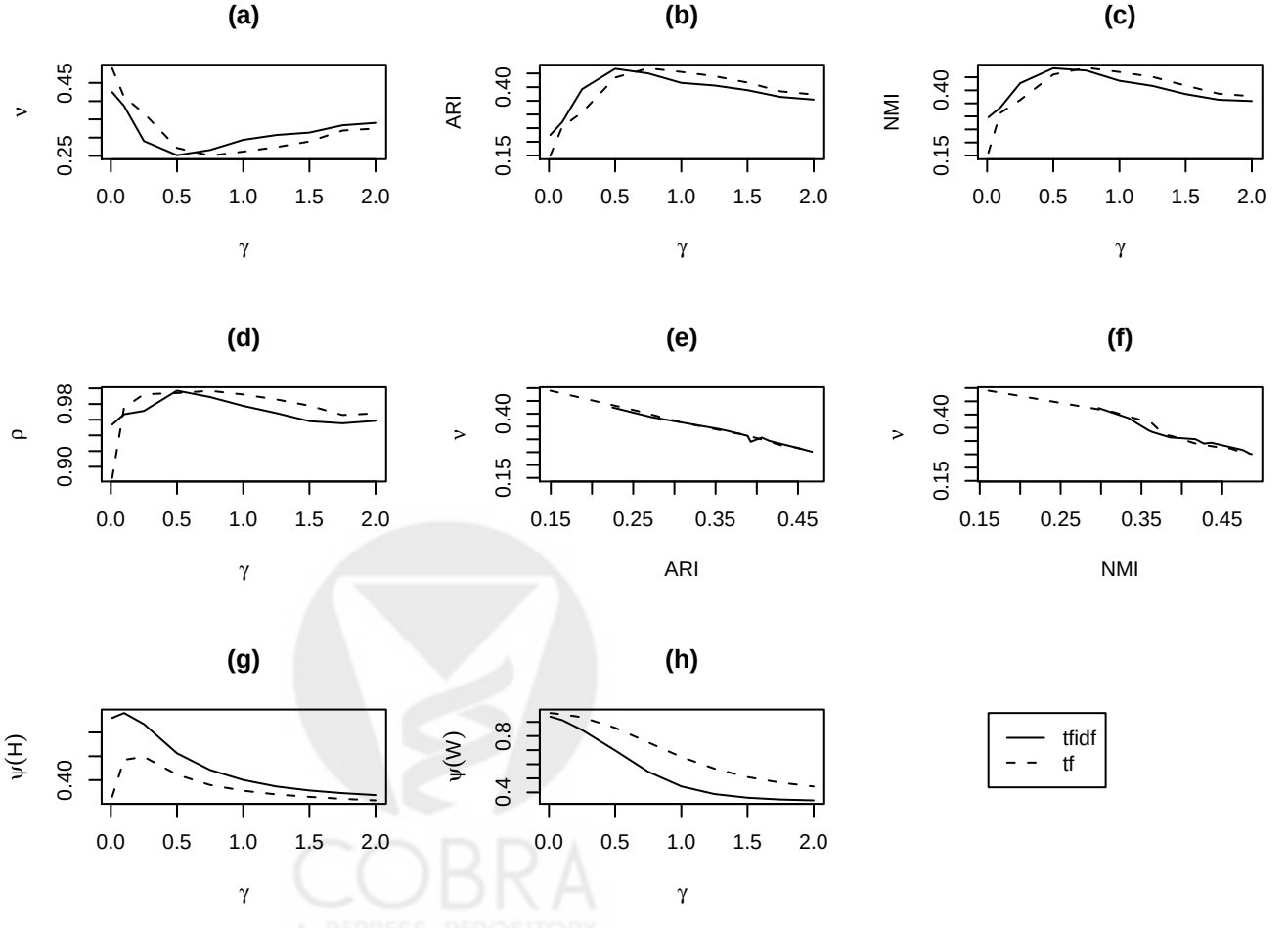


Figure 2: Reuters Data

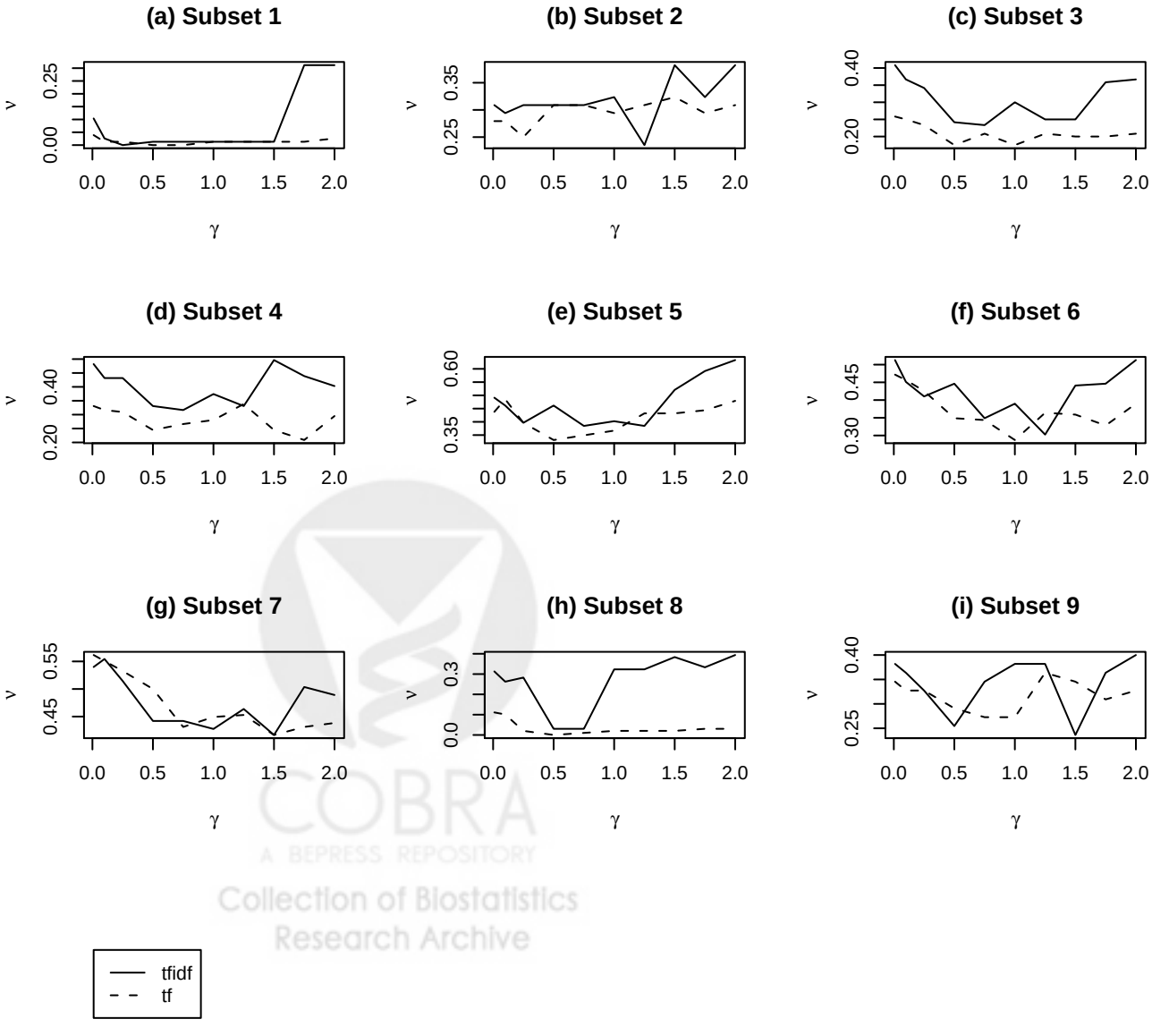
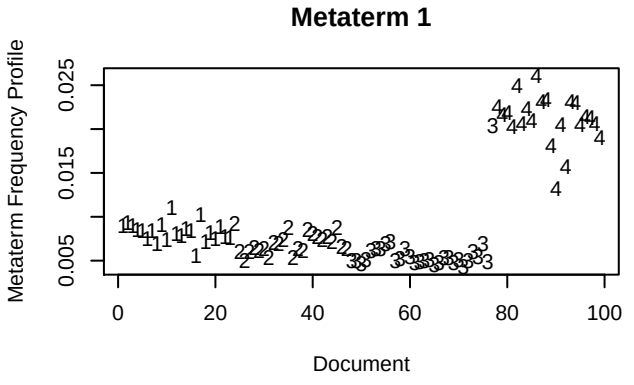
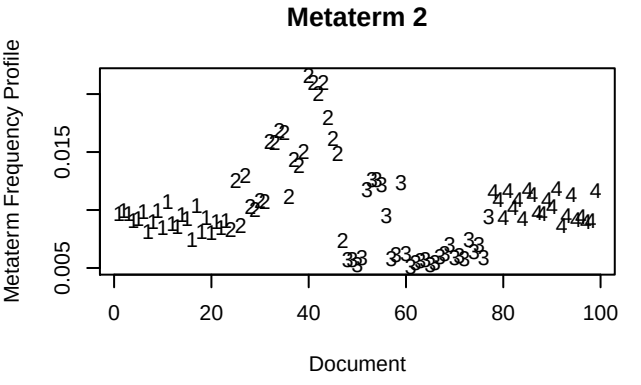
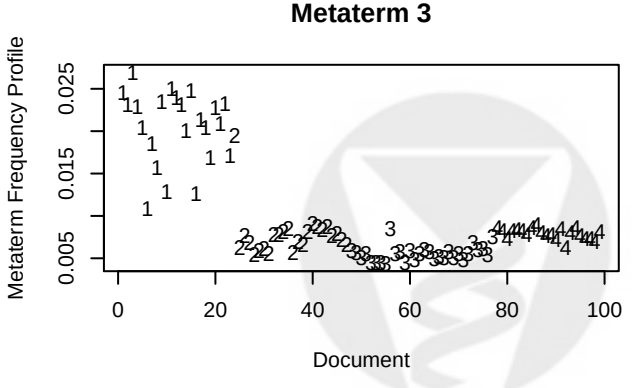


Figure 3: Metaterm Frequency Profiles







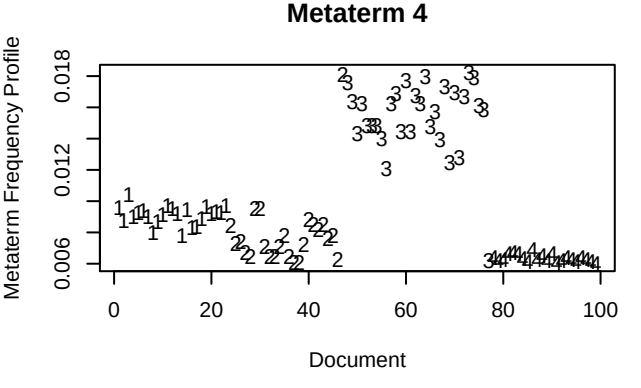


Figure 4: Page Blocks Data

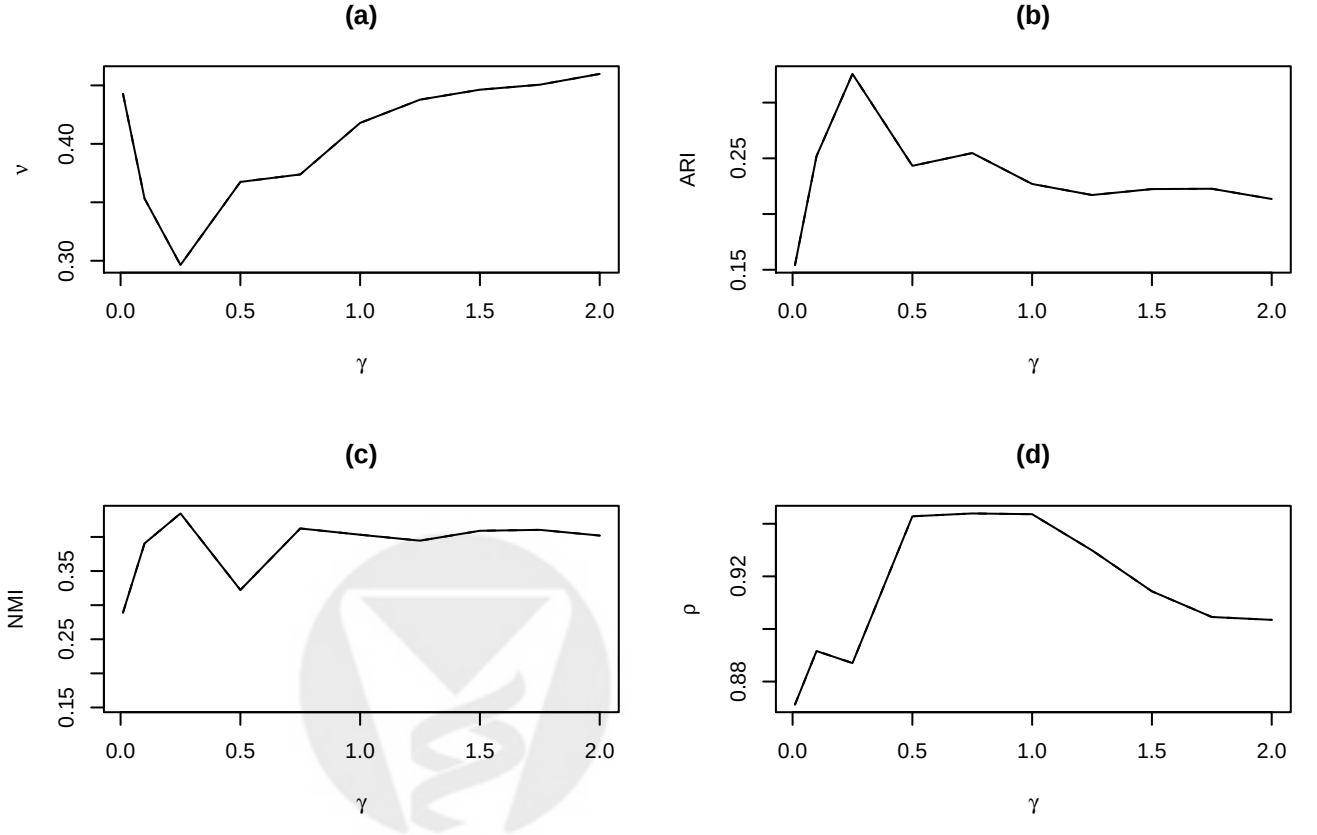


Figure 5: Simulated Data

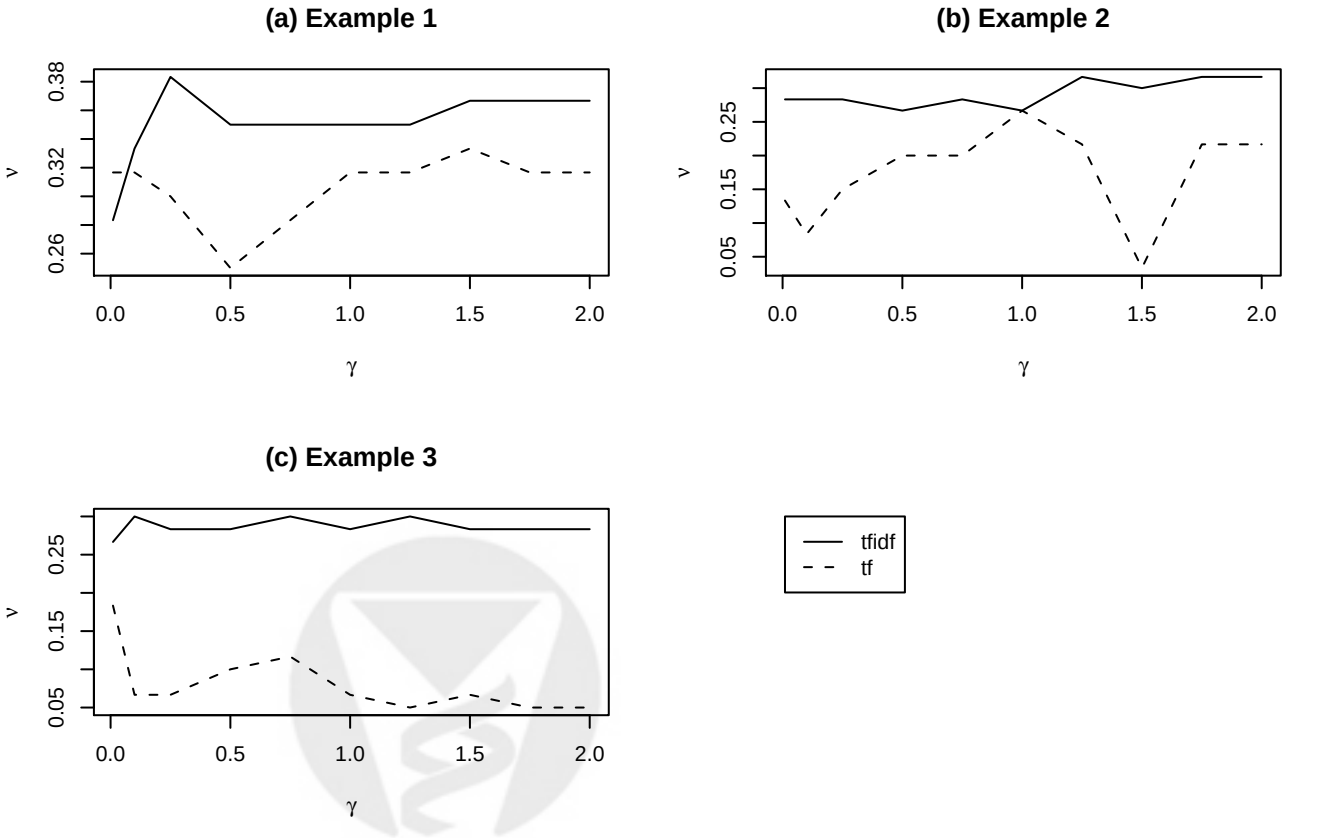
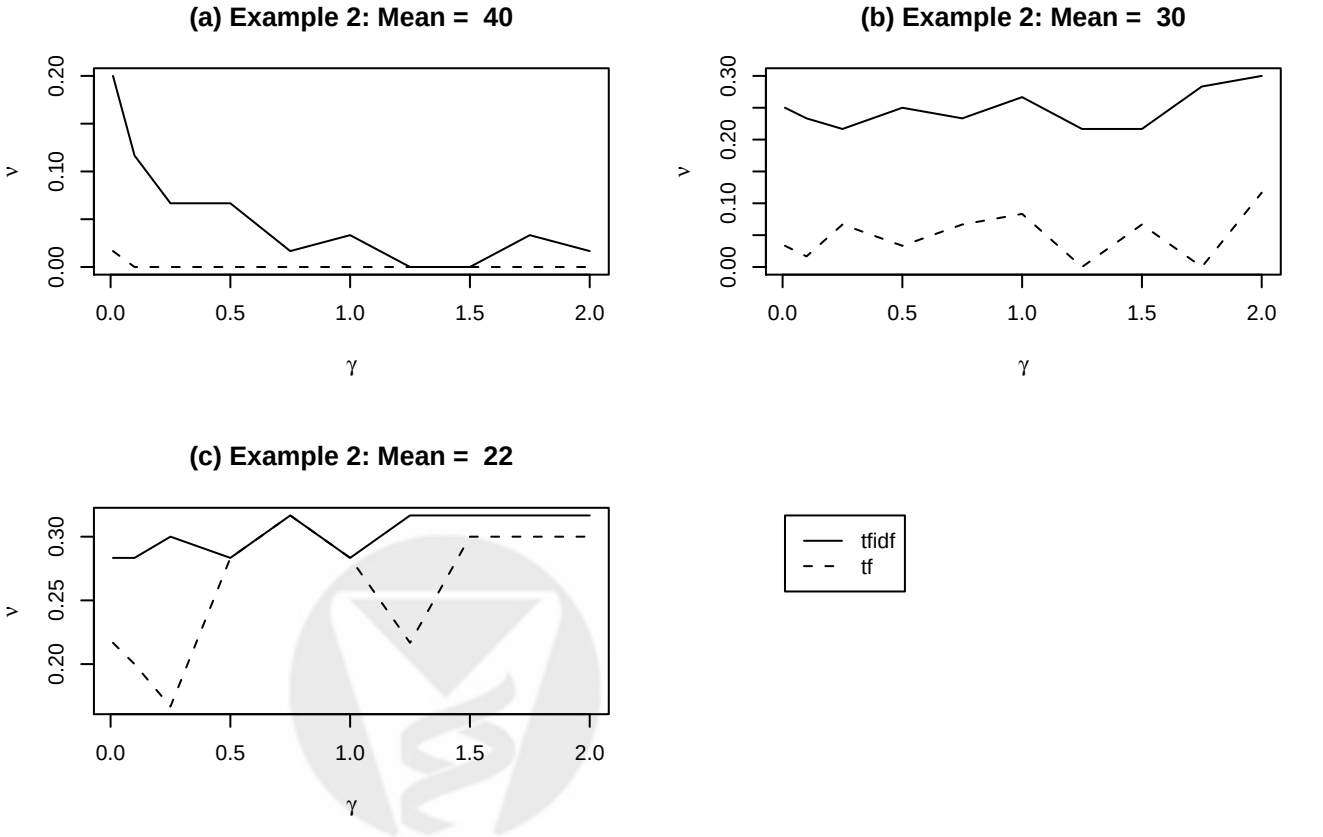
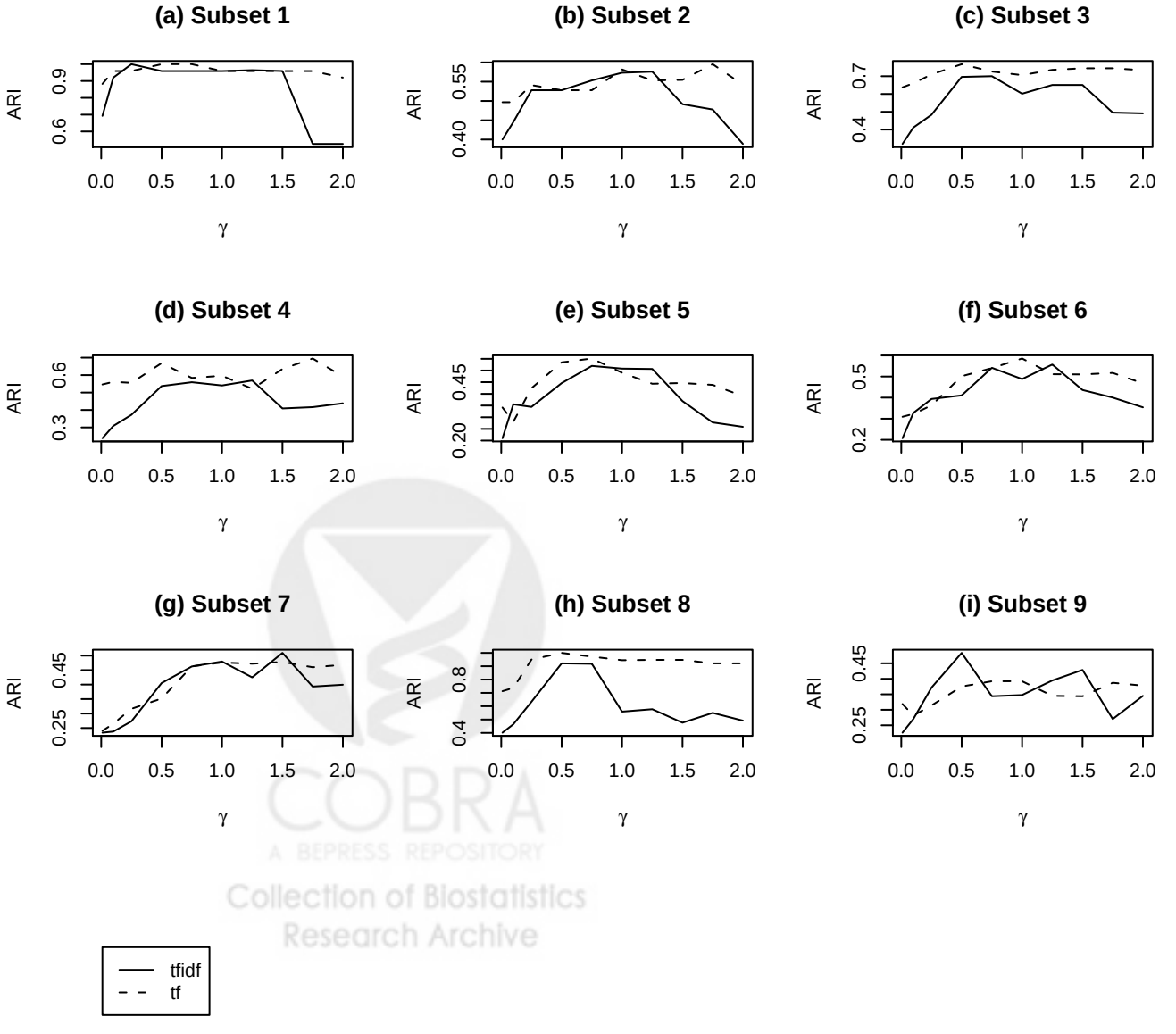


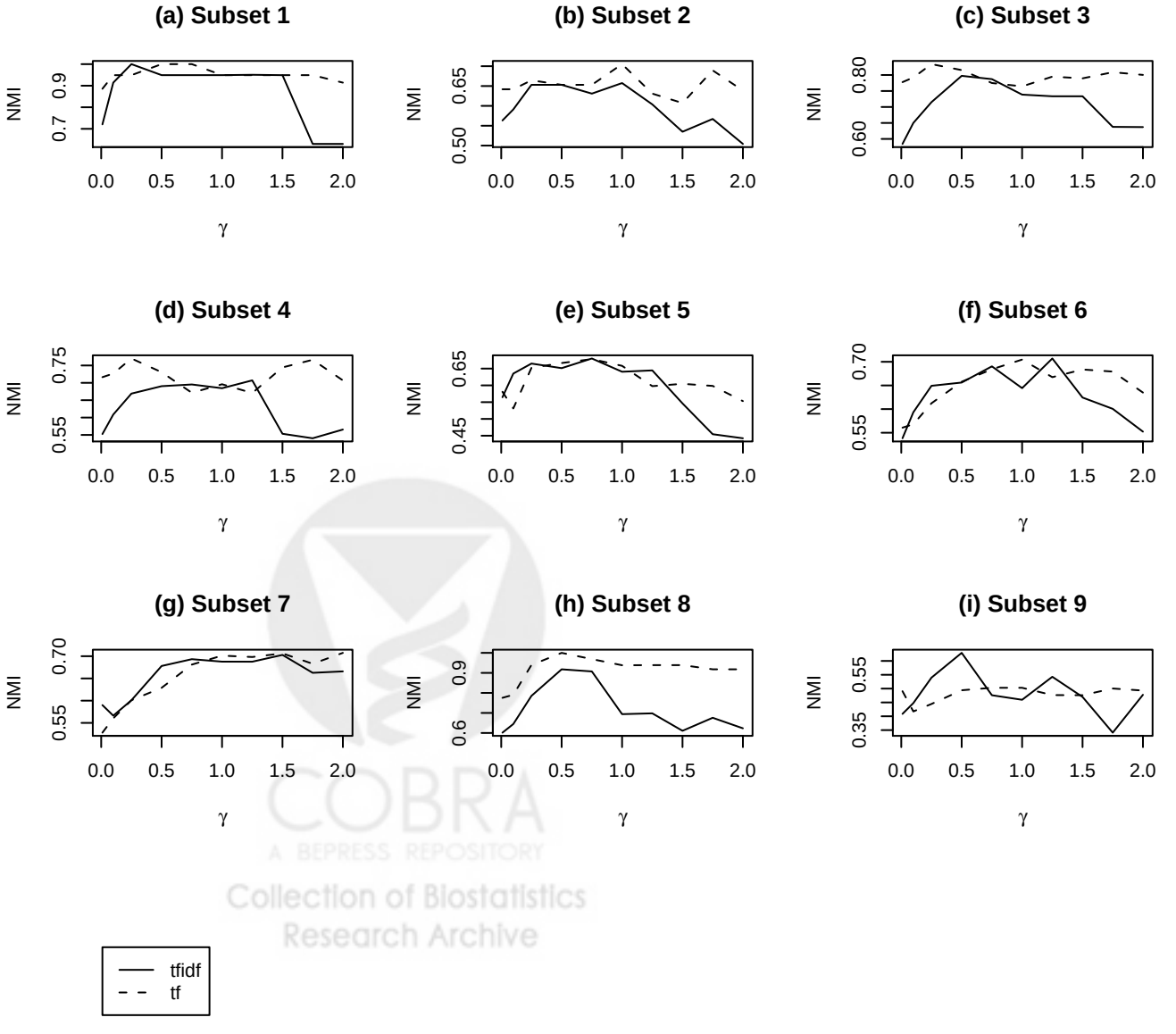
Figure 6: Simulated Data



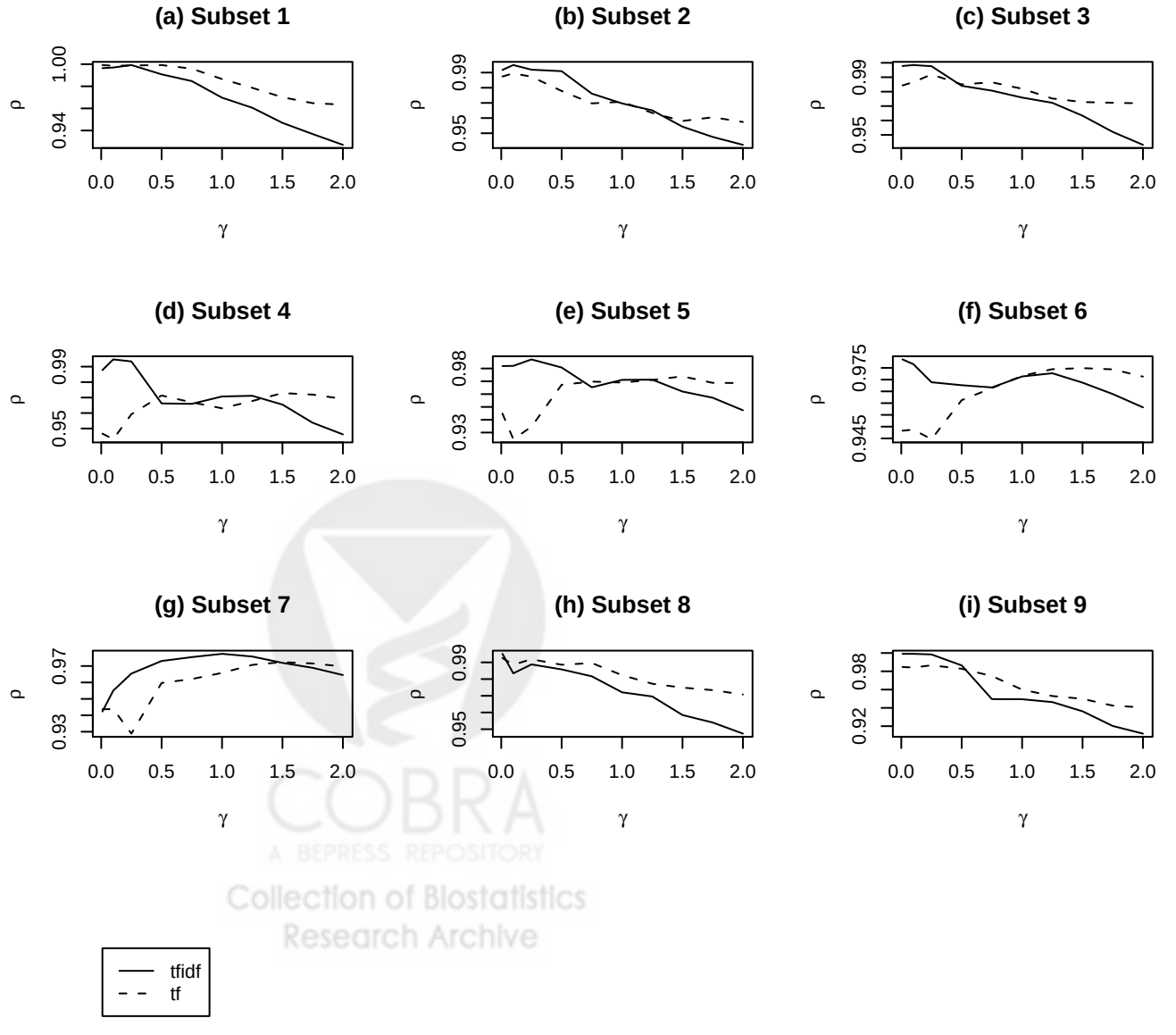
Supplemental Figure S1: Reuters Data



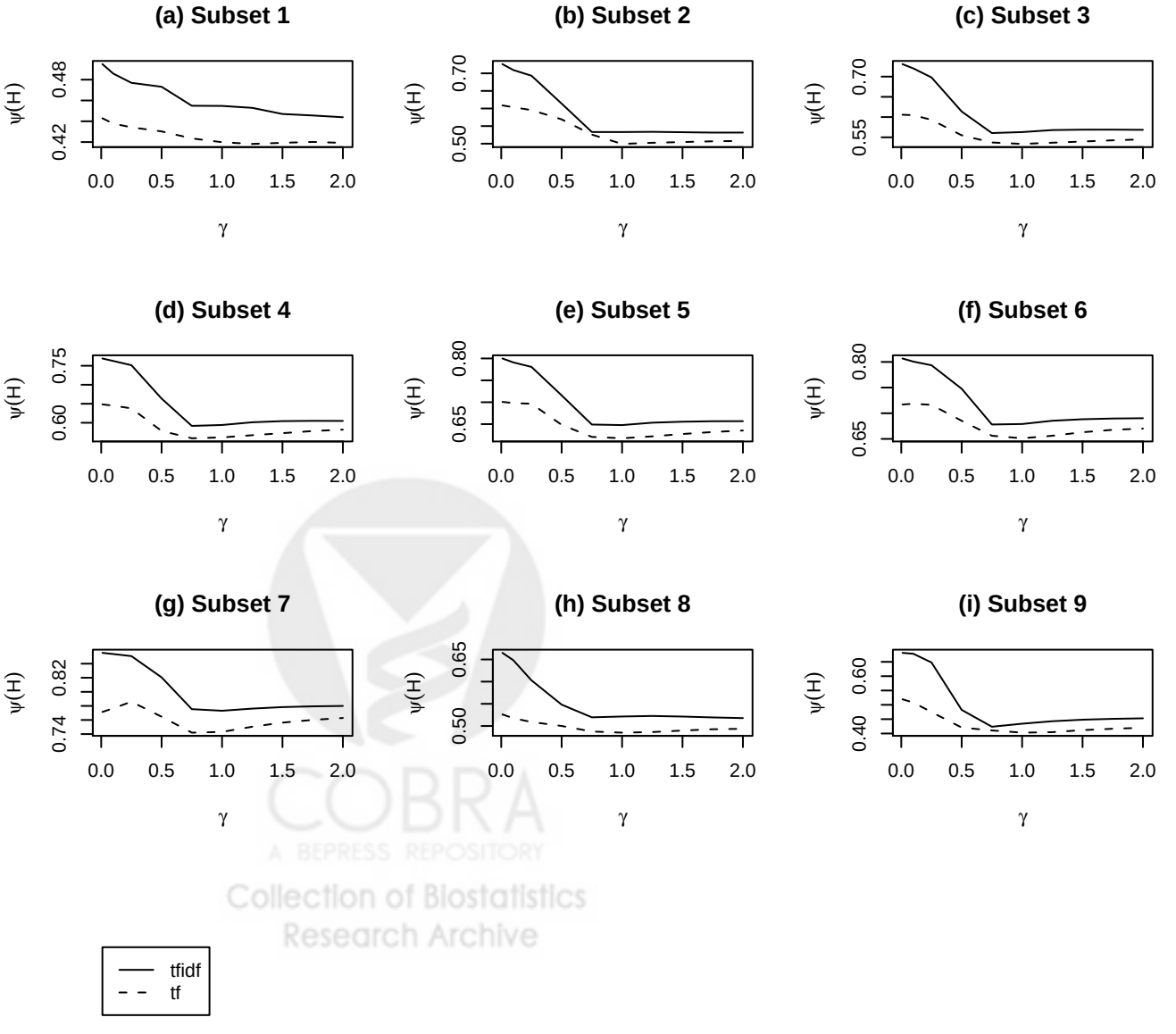
Supplemental Figure S2: Reuters Data



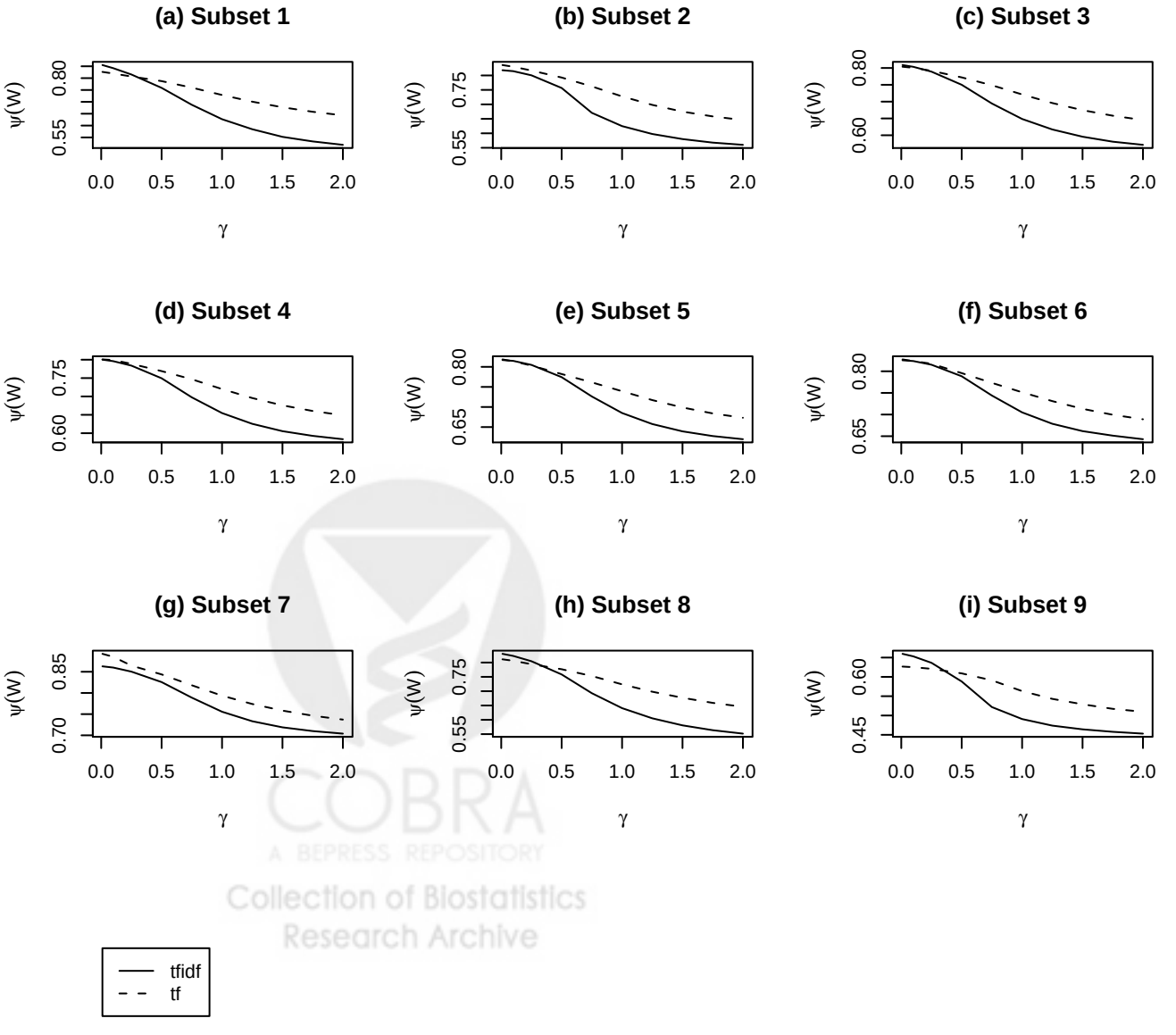
Supplemental Figure S3: Reuters Data



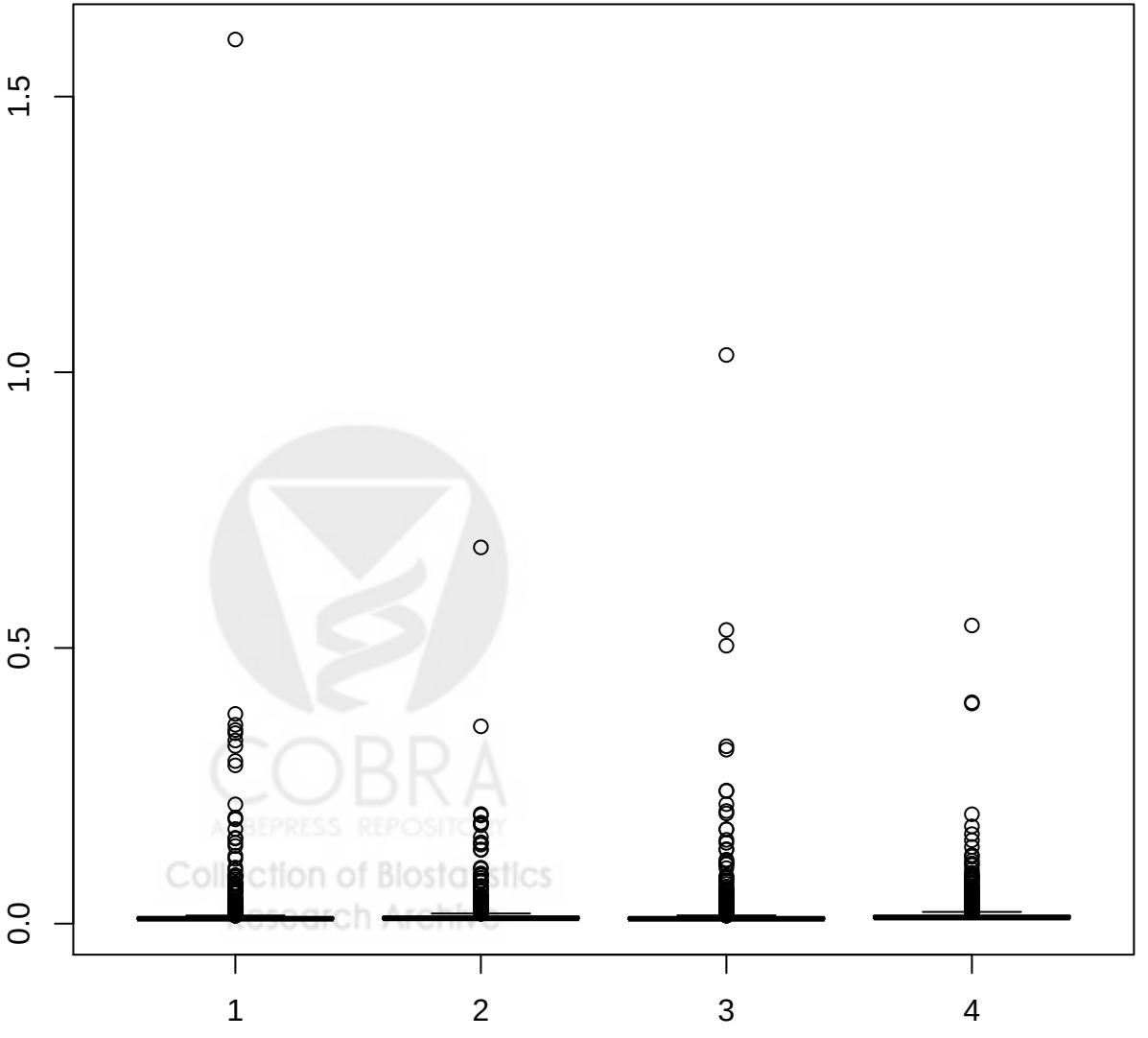
Supplemental Figure S4: Reuters Data



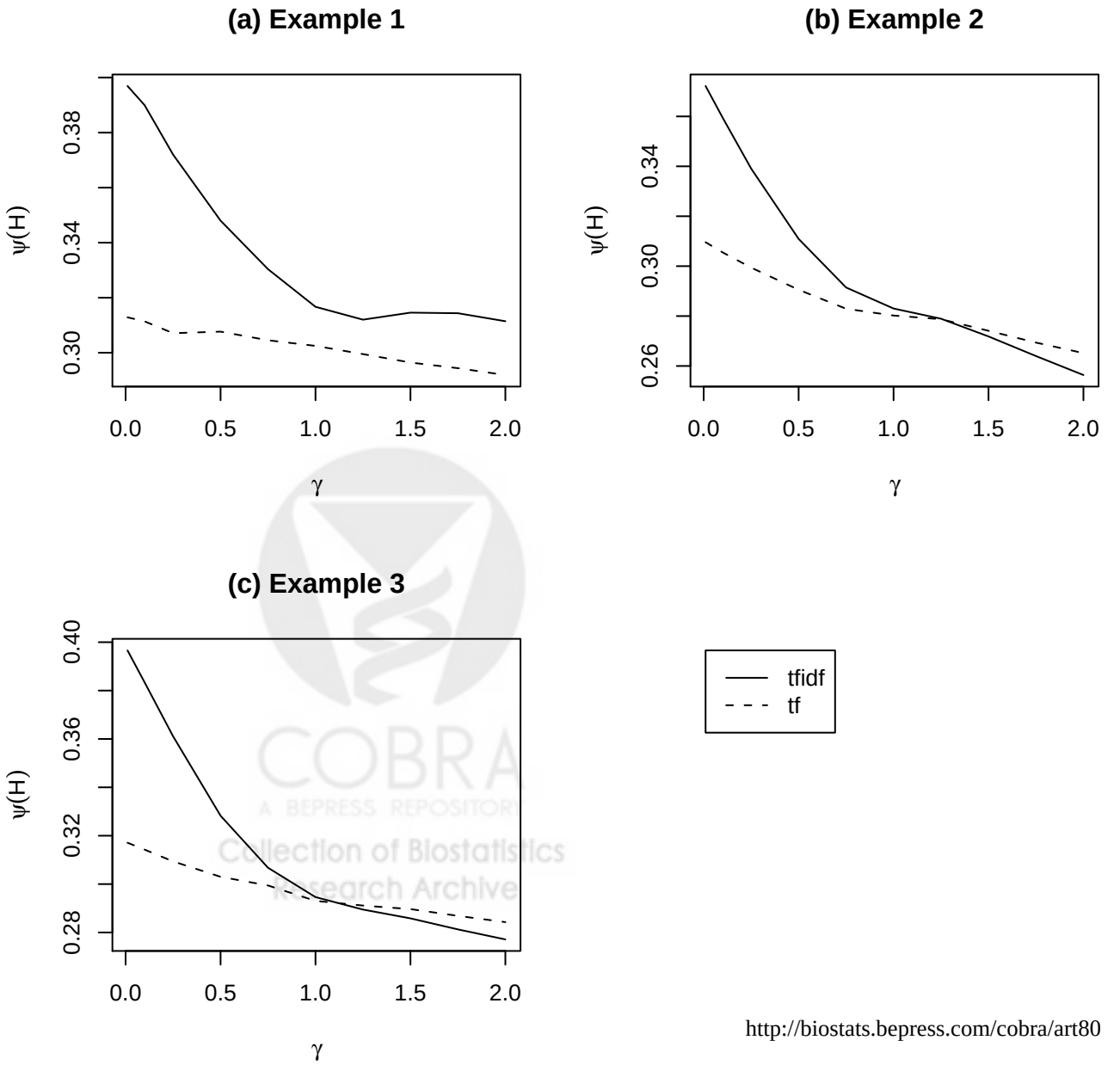
Supplemental Figure S5: Reuters Data



Supplemental Figure S6: Distribution of Metaterms



Supplemental Figure S7: Simulated Data



Supplemental Figure S8: Simulated Data